Physics of the Earth

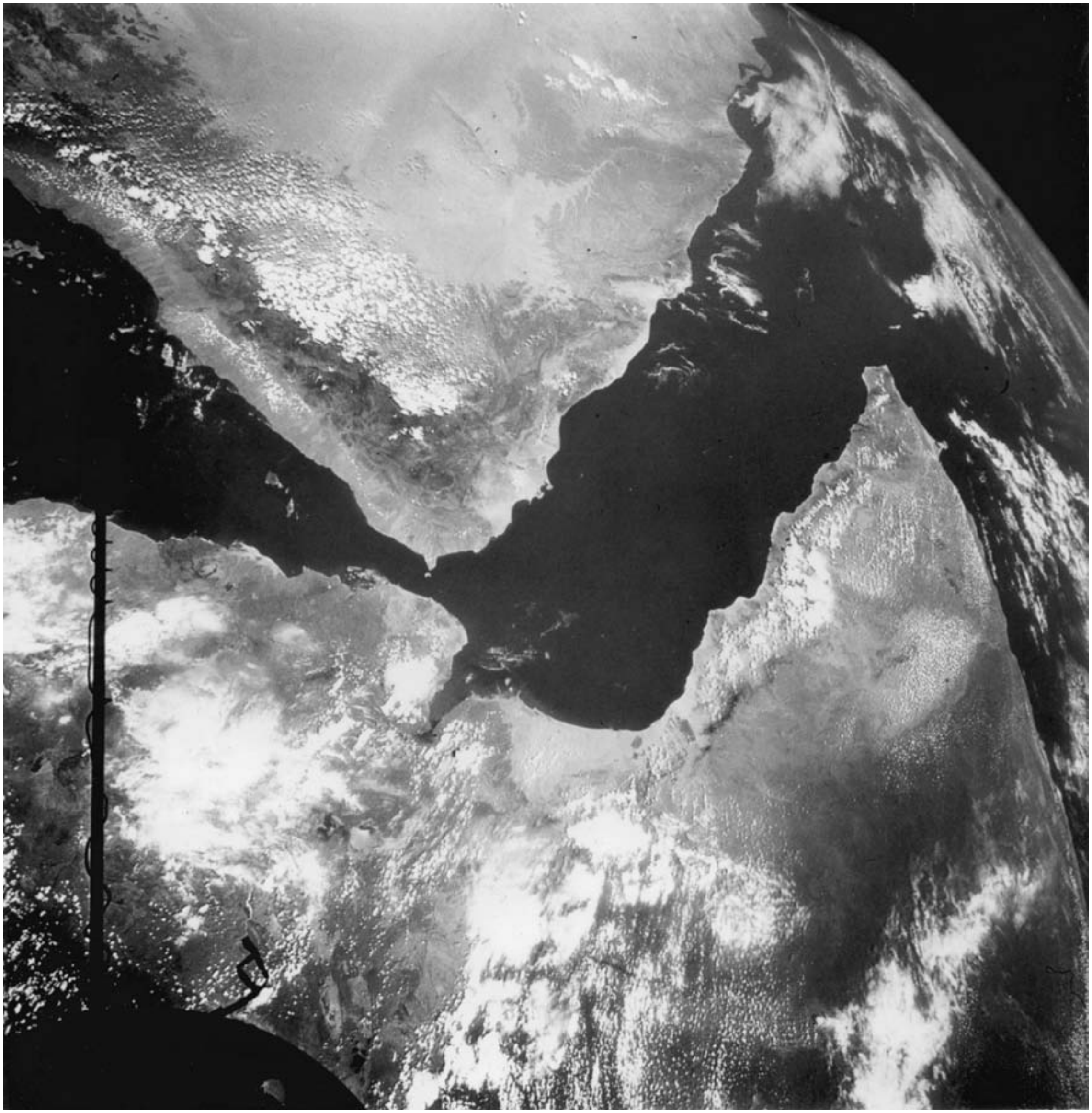
The fourth edition of Physics of the Earth maintains the original philosophy of this classic textbook on fundamental solid Earth geophysics, while being completely revised and up-dated by Frank Stacey and his new co-author Paul Davis. Building on the success of previous editions, which have served generations of graduate students and researchers for nearly forty years, this new edition will be an invaluable resource for graduate students looking for the necessary physical and mathematical foundations to embark on their own research careers in geophysics.

The book presents a detailed, critical analysis of the whole range of global geophysics topics and traces our understanding of the Earth, from its origin and composition to recent ideas about rotation of the inner core. The division of this new edition into an increased number of shorter chapters is designed to make the material more accessible, and allows students to focus on topics of particular interest. New chapters on elastic and inelastic properties, rock mechanics, kinematics of earthquake processes, earthquake dynamics and thermal properties have been added. A brief concluding chapter also reviews contributions from solid Earth studies to our understanding of climate change and the potential for 'alternative' energies.

Appendices, presenting fundamental data and advanced mathematical concepts, and an extensive reference list, are provided as tools to aid readers wishing to pursue topics beyond the level of the book. Over 140 student exercises of varying levels of difficulty are also included, and full solutions are available online at www.cambridge.org/9780521873628.

Frank Stacey is a graduate of London University. After appointments in Canada, Australia and UK, he went to the University of Queensland in 1964 and it was there that the first three editions of 'Physics of the Earth' were written. After retirement as Professor of Applied Physics, he joined CSIRO Exploration and Mining (in 1997) to continue geophysical research. He has published on a wide range of geophysical topics and has been recognized by his peers by election to fellowship of the Australian Academy of Science and the American Geophysical Union and by the award of the inaugural Neel medal of the European Geophysical Society, as well as numerous visiting lectureships at institutions around the world. Professor Stacey is also the author/editor of three other books.

Paul Davis is a graduate of the University of Queensland. After appointments in Edmonton, Canada, and Cambridge, he joined the University of California at Los Angeles (UCLA), where he is Professor of Geophysics. He has published extensively on geophysical topics, especially seismology. His professional honours include a Guggenheim fellowship, fellowship of the Royal Astronomical Society and the American Geophysical Union and a visiting Leverhulme professorship to the University of Oxford. He has served a term as editor of the Journal of Geophysical Research (Solid Earth). Professor Davis is also the co-author of another undergraduate textbook.



GEMINI XI photograph of the Gulf of Aden and the Red Sea by NASA astronauts Charles Conrad and Richard F. Gordon. This is one of the areas of particular interest in the theory of sea floor spreading. A line of earthquake epicentres extends from the ridge system in the Indian Ocean, up the middle of the Gulf of Aden and into the Red Sea, marking the axis of a new ridge along which mantle material is rising as the Africa and Arabia plates part. Courtesy of the National Aeronautics and Space Administration, Washington.

Physics of the Earth

Fourth edition

Frank D Stacey
CSIRO Exploration and Mining, Brisbane, Australia

Paul M Davis
Department of Earth and Space Sciences, University of California, Los Angeles, USA

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo, Delhi

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

Information on this title www.cambridge.org/9780521873628

First published 2008

Printed in the United Kingdom at the University Press, Cambridge

A catalogue record for this publication is available from the British Library

Library of Congress Cataloguing in Publication data

ISBN 978-0-521-87362-8 hardback

Contents

Preface

As with previous editions of this title, our principal aim is to present a coherent account of the Earth that will satisfy advanced students with diverse backgrounds. We have endeavoured to explore the physical principles of the subject in a way that encourages critical appraisal. This requires the reader to have some familiarity with a wide range of inter-related ideas, for which there is no clearly preferred, logical order of presentation. Should the properties of meteorites precede or follow the isotopic methods used to study them? Is it important to understand something about the Earth's internal heat before studying seismology or vice versa? Can we be clear about the evidence for tectonic activity without knowing about the behaviour of the geomagnetic field? We have attempted to avoid the need for answers to these questions by beginning each chapter with what we call a preamble. Our preambles are not intended to be synopses of the chapters or even introductions in the conventional sense, but glue to hold the subject together, with glimpses of related concepts from other chapters. We hope to convey in this way a feel for the unity of the subject. Especially for students using this book as a text, we suggest reading all of the preambles before looking deeper into any of the chapters.

The appendices and the list of references are also indications of our philosophy. They are included as tools to aid students, or others, who are pursuing topics beyond the level of this book, questioning the approach we have taken or simply seeking convenient reference material. We often learn most effectively by doubting something we read and conducting an independent check, either by a calculation or by a literature search. This is especially true in using a text such as ours, which introduces ideas that are recent and await confirmation or are even disputed. One of the appendices is a set of problems, many of which we have used with our own classes. They have a wide range of sophistication, from near trivial to difficult. For convenience they are numbered to identify them with particular chapters, but in many cases it is not clear to which chapters they are most relevant. Problems that provide bridges between topics are probably the most useful and we draw attention to some of them in the text. Our own solutions are presented on a website: www.cambridge.org/9780521873628.

We like to think that this book will be read by the next generation of geophysicists, who will develop an understanding of things that currently puzzle us or correct things that we have got wrong. We refer in the text to some of the tantalizing questions that await their attention and they will find more that we have not thought of. Advice about our errors, omissions and obscurities will be appreciated. We thank colleagues who have reviewed draft chapters and helped us to minimize the flaws: Charles Barton, Peter Bird, Emily Brodsky, Shamita Das, David Dunlop, Emily Foote, Mark Harrison, Donald Isaak, Ian Jackson, Mark Jacobson, Brian Kennett, Andrew King, Frank Kyte, David Loper, Kevin McKeegan, Ronald Merrill, Francis Nimmo, Richard Peltier, Henry Pollack, Joy Stacey, Sabine Stanley and George Williams.

Frank Stacey
Paul Davis

1

Origin and history of the Solar System

1.1 Preamble

As early astronomers recognized, the planets are orbiting the Sun on paths that are nearly circular and coplanar, with motions in the same sense as the Sun's rotation, making it difficult to avoid the conclusion that the formation of the Sun led directly to its planetary system. A comparison of the planets (Table 1.1) shows that the Earth belongs to an inner group of four, which are much smaller and denser than the outer four giant planets. For this reason the inner four are referred to as the terrestrial (Earth-like) planets. The outer four large planets are gaseous, at least in their visible outer regions, but they have solid satellites with very varied external appearances and a wide range of mean densities, all much lower than those of the terrestrial planets. The asteroids, which are found between the two groups of planets, are believed to be remaining examples of planetesimals, the pre-planetary bodies from which the terrestrial planets formed. Meteorites are samples of asteroids that have arrived on the Earth, by a mechanism discussed in Section 1.9, and so provide direct evidence of the overall compositions of the terrestrial planets.

We probably learn more about the formation of planets from their differences than from their similarities. Several features of the Earth distinguish it from all other bodies in the Solar System and require special explanations.

(i) The Earth is the only planet with abundant surface water, both liquid and solid.

(ii) It is also the only planet with an atmosphere rich in oxygen.

(iii) It appears to be the only planet with extensive areas of acid, silica-rich rocks, such as granite, which are characteristic of the Earth's crust in continental areas.

(iv) It is usually regarded as the only planet with a bimodal distribution of surface elevations (Fig. 9.4), marking the division into continental and oceanic areas, although it is possible that Mars has weak evidence of a similar crustal structure (Section 1.14).

(v) The Earth is the only terrestrial planet with a strong magnetic field. In this respect it resembles the giant planets (Table 24.2).

(vi) Perhaps the existence of a large moon should be added to the list of features that make the Earth unique, at least among the terrestrial planets. The origin and history of the Moon are the subjects of rival ideas. Sections 1.15 and 8.6 address this problem.

The first four of these features are related and water provides the connecting link. It is necessary for the plant life that has produced the atmospheric oxygen. It is also essential to the tectonic process that leads to acid volcanism (Section 2.12). The acid rocks that form the basis of the continents are lighter than the underlying mantle and 'float' higher in the gravitational (isostatic) balance of the Earth's crust (Section 9.3), causing the bimodal distribution of surface elevations.

Meteorites are especially important to our understanding of the early history of the Solar

Table 1.1 Planetary parameters

n Planet	Orbit radius[a] (AU) (r_n/r_3)	Rotation period (days)	Tilt of equator to orbit (degrees)	Mass (Earth masses)	Radius[b] (Earth radii)	Mean density (kg m^{-3})	Decompressed density (kg m^{-3})	Number of known satellites
1 Mercury	0.387	59	2.0	0.055 28	0.3830	5427	5017	0
2 Venus	0.723	243.02	177.3	0.814 999	0.9499	5204	3868	0
3 Earth	1	0.997 2697	23.45	1	1	5515	3995	1
Moon				0.012 3000	0.2728	3345	3269	
Earth + Moon				1.012 3000			3945	
4 Mars	1.524	1.026	25.2	0.107 4468	0.5321	3933	3697	2
[5] Asteroids	~2.8					3700[c]	3700	
6 Jupiter	5.2013	0.413	3.1	317.89	10.973	1327	largely	62
7 Saturn	9.538	0.444	26.7	95.18	9.140	688	gaseous	35
8 Uranus	19.18	0.718	97.9	14.54	3.98	1272		27
9 Neptune	30.06	0.671	30.2	17.15	3.86	1640		13
10 Pluto	39.52	6.387	117.6	0.0022	~0.18	2080	~2000	3

[a] Semi-major axis of orbital ellipse
[b] Radius of a sphere of equal volume (for surface at 1 atmosphere pressure for outer planets)
[c] Average of observed falls

System (Sections 1.7 to 1.11) and its composition (Sections 2.2 to 2.5). Most of them are fragments of asteroids. They are samples of small bodies with relatively simple histories that have remained virtually unaltered since the Solar System was formed. Collisions in the asteroidal belt projected these fragments into orbits that evolved, initially by the Yarkovsky effect (Section 1.9) and then by orbital resonances with Jupiter, into Earth-crossing paths, providing us with samples that are more representative of the total chemistry of the terrestrial planets than is the Earth's crust. For a broad review of their importance to our understanding of early Solar System history see Wasson (1985).

Our estimate of the age of the Solar System is derived from measurements of isotopes produced in meteorites by radioactive decay (Section 4.3). It is clear that most of them have a common age, 4.57×10^9 years. The evidence that this dates the formation of the whole Solar System is less direct because we do not find on the Earth any rocks that have survived that long unaltered. However, by obtaining an estimate of the average isotopic composition of the Earth as a whole we can plot it as a single data point on the isochron (equal-time line) of meteorite lead isotopes (Fig. 4.1) and see that it fits reasonably well.

1.2 Planetary orbits: the Titius–Bode law

For many years theories of the origin of the Solar System were based on rather little hard evidence. Motions of the planets were well observed and orbital radii were seen to follow a regular, if approximate, pattern. This regularity was represented by an equation known as the Titius–Bode law or sometimes simply as Bode's law. As originally proposed this law gave the orbital radius of the kth planet (counted outwards) as

$$r_k = a + b \times 2^k, \tag{1.1}$$

a and b being constants. Modern discussions favour a power law but still refer to it as the Titius–Bode law,

$$r_k = r_0 p^k. \tag{1.2}$$

Although we now have much more information about the planets, this relationship is still central to our understanding of the Solar System.

The choice of value of p in Eq. (1.2) depends on how the planets are counted. The wide gap between Mars and Jupiter led to the search for a 'missing' planet in the region now recognized to be occupied only by numerous asteroids. We no longer suppose that the asteroids were ever parts of one or two planets, and cannot logically fit the Titius–Bode law to the whole set of planets. Attempts to do this (the broken line in Fig. 1.1) are only of historical interest and we should fit Eq. (1.2) to the two groups of planets separately (the solid lines in the figure). Pluto must not be considered with the outer group as it is identified with the pre-planetary fragments (Kuiper belt objects, Section 1.13) that have escaped accretion into the regular planets. But, in spite of these difficulties, the approximate geometrical progression of orbital radii is clear and obviously has a fundamental cause. As evidence of the generality of the Titius–Bode law, we note that the orbital radii of the major satellites of the giant planets also fit Eq. (1.2). The fit is particularly good for the satellites of Jupiter (Io, Europa, Ganymede, Callisto) which give $p = 1.64 \pm 0.03$, in the same range as for the planetary fit. Bode's law is an interesting example of scale invariance, which is seen in many physical phenomena. It is an apparently universal law, independent of the scale of the orbits and of the masses of planets or satellites.

Theories to explain the Titius–Bode law abound, as discussed in a historical review by Nieto (1972). There are two basic approaches, appealing to regularities in the scale of turbulence in the solar nebula or to the competition between gravitational attractions of aggregating bodies in the gradient of the solar gravity field. The vortex theories were pioneered by Laplace and more recently by R. P. von Weizsäcker. They were given a focus by White (1972), who argued that the nebular cloud was sufficiently tenuous to have negligible viscosity, so that vorticity was conserved. With this assumption, White's theory leads to jet streams spaced radially from the Sun in the manner of Eq. (1.2). Prentice (1986, 1989)

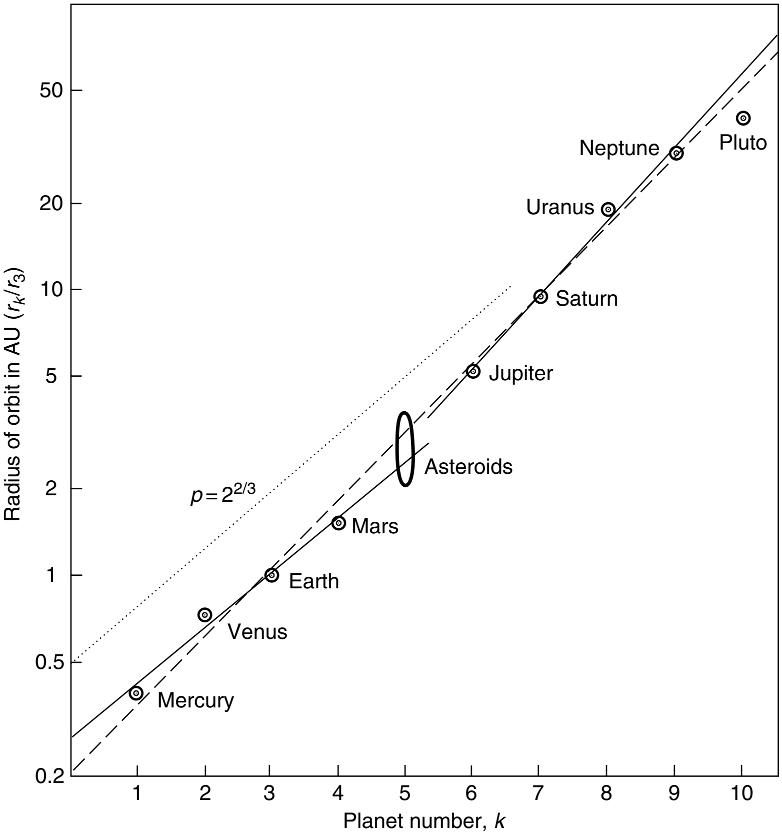


FIGURE 1.1 Radii of planetary orbits fitted to the Titius–Bode law (Eq. 1.2). Solid lines are independent fits to the terrestrial planets ($p = 1.56$) and the four giant planets ($p = 1.82$). The broken line is a fit to all planets except Pluto, allowing for a missing fifth (asteroidal) planet ($p = 1.73$). The dotted line indicates a gradient for $p = 2^{2/3} = 1.59$, which would apply if all orbital periods of neighbouring planets differed by a factor of 2 (Kepler's third law – Eq. (B.23) in Appendix B).

pointed out that the turbulence would have been highly supersonic, and developed a theory of planetary accretion on that basis. White's point about the low viscosity draws attention to the need for a mechanism to damp the turbulence. In Section 4.6 we argue that that the only plausible one is electromagnetic and could not have operated until the arrival of highly radioactive supernova debris ionized the nebular material and made it sufficiently conducting for hydromagnetic damping to occur.

There can be little doubt that mutual gravitational attraction of planetesimals had a role in planetary accretion, at least in its late stage, and there are several variants of the gravitational interpretation of Bode's law. The simple observation, that the parameter p in Eq. (1.2) is very close to $2^{2/3} = 1.587$, invites close scrutiny. By Kepler's third law (Eq. B23 in Appendix B), this ratio of orbital radii corresponds to orbital periods with 2:1 ratios. It is the simplest of the resonances that have been intensively studied in connection with asteroids, but, in that case, interaction with Jupiter is of interest rather than mutual interactions between small bodies. Attempts to explain Bode's law in terms of gravitational interactions appeal to the fact that orbital speeds of planetesimals decreased as $a^{-1/2}$ with distance, a, from the Sun. Mutual interactions extended over ranges that increased systematically with a by virtue of the decreasing differential speeds. An unambiguous theory of Bode's law eludes us, but the evidence for universal validity of Eq. (1.2) is compelling. We apply it as an empirical rule in discussing the early history of the Moon in Section 8.6.

1.3 Axial rotations

The rotations of the planets differ greatly from one another in both speed and axial orientation

(Table 1.1). Rotation in the sense of the orbital motion predominates, but Uranus, whose axis is almost in the orbital plane, and Venus, whose very slow rotation is retrograde, are exceptions. The conventional explanation for the variation in axial alignment is the same statistical one as is offered for the scatter of orbital radii about a regular pattern (Fig. 1.1), that is, it depends on the infall of planetesimals on independent, but reasonably close orbits. Precisely how they collided determined the rotations of the composite bodies.

The terrestrial planets and Pluto are rotating slowly compared with both the giant planets and the asteroids. In the cases of Mercury, Venus, the Earth and Pluto, the slower rotations are due to the dissipation of rotational energy by tidal friction (discussed in Section 8.3). The rotation of Mercury is believed to be tidally locked to its elliptical orbit about the Sun. The tide raised in Pluto by its satellite, Charon, has stopped it rotating relative to Charon, to which it presents a fixed face, as does the Moon to the Earth for the same reason. Venus must have been slowed by friction of the solar tide, but that does not explain the retrograde sense of its rotation, which must therefore be a consequence of the accretion process. Mars is too far from the Sun and its present satellites are too small for tidal friction to have had a noticeable effect on it and, unless it once had a large, close satellite that spiralled in and merged with the planet (as suggested for Mercury and Venus in Section 1.15), the slow rotation is what it was left with after the arrival of the last planetesimal. The near coincidences of the rotational speeds and axial alignments (obliquities) of Mars and the Earth are fortuitous. The early rotation of the Earth was certainly much faster and the obliquity of Mars is subject to variation by gravitational interactions, especially with Jupiter.

The rotation of Uranus, with its axis close to the orbital plane, gives a clue to the mechanism of planetary accretion. The silicate and iron content could not reasonably be sufficient to account for the rotational angular momentum, even if it arrived as a tangentially incident planetesimal. We therefore suppose that Uranus accreted from volatile-rich planetesimals that had formed in independent solar orbits. Direct accretion from gas could not have caused the axial misalignment if dissipation of turbulence in the nebula and collapse to a disc preceded planetary formation. This would have confined the motion of the gas more or less to the plane of the disc, from which random accretion of very large numbers of molecules could not have resulted in planetary rotation perpendicular to the plane. The planetesimals would have been composed of ices that could condense out of the nebula, and not hydrogen or helium, which could have accreted only on a planet that was large enough to hold them gravitationally. In the case of Jupiter, for which hydrogen and helium represent a much larger proportion of the total mass, the axial misalignment is very slight.

1.4 Distribution of angular momentum

Using Kepler's third law (Eq. B.23, Appendix B) we can write the orbital angular velocity of the *k*th planet,

$$\omega_k = (GM_S/r_k^3)^{1/2}, \tag{1.3}$$

in terms of its orbital radius r_k, the mass of the Sun, $M_S = 1.989 \times 10^{30}$ kg and the gravitational constant, G. This allows us to write the orbital angular momentum in a convenient form,

$$a_k = m_k r_k^2 \omega_k = (GM_S)^{1/2} m_k r_k^{1/2}, \tag{1.4}$$

for calculation of the total orbital angular momentum of the Solar System from Table 1.1,

$$\begin{aligned}\sum a_k &= (GM_S)^{1/2} \sum m_k r_k^{1/2} \\ &= 3.137 \times 10^{43}\ \text{kg m}^2\text{s}^{-1},\end{aligned} \tag{1.5}$$

Jupiter accounts for more than 60% of this total.

The angular momenta of planetary rotations are very much smaller than the orbital angular momenta. The rotational angular momentum of the Earth, $5.860 \times 10^{33}\,\text{kg}\,\text{m}^2\,\text{s}^{-1}$, is 2.2 parts in 10^7 of its orbital angular momentum, $2.662 \times 10^{40}\,\text{kg}\,\text{m}^2\,\text{s}^{-1}$. We can compare Eq. (1.5) with the rotational angular momentum of the Sun, which has 99.866% of the total mass of the Solar System (assuming that we know about it all). The surface of the Sun is rotating faster in equatorial regions than at the poles and, although there is no direct observation to indicate how the interior is rotating, observations of the modes of free oscillation (helioseismology) are consistent with coherent rotation, so it suffices for the present purpose to assume rigid body rotation with the angular speed taken as representative by Allen (1973), $\omega_S = 2.865 \times 10^{-6}\,\mathrm{rad\,s^{-1}}$. The rigid body moment of inertia can be obtained by integrating the density profile of the Sun (Problem 1.3, Appendix J), which has a strong concentration of mass towards the centre. Allen's (1973) value is $5.7 \times 10^{46}\,\mathrm{kg\ m^2}$. With the above value of ω_S, this gives the angular momentum

$$(I\omega)_S = 1.63 \times 10^{41}\,\mathrm{kg\,m^2\,s^{-1}}. \tag{1.6}$$

Thus the Sun has only a small fraction (0.5%) of the angular momentum of the Solar System, which is dominated by the planetary orbits (Eq. 1.5), although the planets have little more than 0.1% of the mass.

The slow solar rotation can be explained by an outward transfer of angular momentum in the nebula which surrounded the Sun when it was still young. Alfvén (1954) argued that this occurred because a strong solar magnetic field (rotating with the Sun) dragged with it the ionized gases of the nebula and an intense solar wind. His suggestion fits well with other observations, especially the magnetizations of meteorites (Section 1.11). Early in the development of the Solar System the Sun is believed to have passed through a stage reached at the present time by a number of young stars (several hundred in our Galaxy), of which T-Tauri is the representative example. They are very active, with strong stellar winds and magnetic fields several orders of magnitude more intense than that of the Sun at present. We suppose that the meteorites were forming when the Sun was at its T-Tauri stage and so were magnetized by its strong field.

The angular momentum transfer by Alfvén's magnetic centrifuge mechanism could have contributed to chemical fractionation in the Solar System. It is only the plasma of charged particles that would be affected by the motion of the magnetic field. Once solid particles began to form, they and any un-ionized gas molecules would have been coupled to the field only by viscous drag of the surrounding plasma. The early condensing, generally less volatile materials would therefore have become relatively more concentrated in the inner part of the Solar System, with most of the volatiles centrifuged to the outer regions.

1.5 Satellites

The giant planets have numerous satellites (Table 1.1), but the terrestrial planets have only three between them and, of these, the Earth's Moon is outstandingly the largest. The other two, Phobos and Deimos, are small, irregularly shaped close satellites of Mars that give the impression of being captured asteroids. The larger one, Phobos, is so close that it orbits Mars three times per day (the Martian and Earth days are almost equal). It has a dark surface, with a reflection spectrum similar to those of many asteroids and to a class of meteorite, the carbonaceous chondrites (Section 2.4). The closeness of the orbit means that Phobos raises an appreciable tide in Mars, in spite of being so small. This makes capture a plausible hypothesis, because it allows the orbit to evolve by tidal friction. Deimos is even smaller and is more remote from Mars, making capture unlikely, although it, too, looks asteroidal.

As well as having many satellites, the giant planets all have rings of fine particles that are most clearly observed around Saturn. In the case of Jupiter, it is apparent that most or all of the small, outer satellites are captured asteroids. Their orbits are tightly clustered in two distinct groups, one prograde at about $11.5 \times 10^6\,\mathrm{km}$ from Jupiter and the other retrograde at about $23 \times 10^6\,\mathrm{km}$. These are the orbits predicted by capture theory. The larger satellites of Jupiter are much closer and, as we mention in Section 1.2, follow the Titius–Bode law. The case for satellite capture by the other giant planets is not as clear, but Neptune's Triton has a retrograde orbit and Nereid a very elliptical one, making capture, or some other vigorous interaction, perhaps with Pluto, appear likely. All the other satellites are presumed to have formed with their parent planets in the same manner as the planets were formed around the Sun. As with the planets, there is a wide range of properties.

Surfaces of the satellites of the giant planets are very different from one another. Extrapolating from our observations of the Moon, we might have expected Voyager images to show ancient, cratered surfaces everywhere. Instead, several satellites show evidence of internal activity and even active volcanism. This is most striking on Jupiter's closest large satellite, Io, where it is attributed to the generation of internal heat by tidal friction (Peale *et al.*, 1979); eccentricity of the close orbit ($e = 0.0043$) is maintained by resonances with other satellites, causing a strong radial tide. Neptune's Triton is another example, and Enceladus, a satellite of Saturn, shows evidence of 'cryovolcanism' of its light ices.

The densities of the satellites of the giant planets are mostly less than 2000 kg m^{-3}, much lower than the densities of terrestrial planets, indicating compositions rich in ices (condensed volatiles such as H_2O, CH_4). The exceptions are Jupiter's innermost two large satellites, Io ($\rho = 3530\,\mathrm{kg\,m^{-3}}$) and Europa ($\rho = 3014\,\mathrm{kg\,m^{-3}}$), which evidently have larger silicate components (and perhaps even small metallic cores). Europa is a case of particular interest. While its surface is permanently frozen hard, a suggestion that it has a liquid ocean at modest depth arises from its influence on Jupiter's magnetic field (Kivelson *et al.*, 2000). Its orbit is within Jupiter's magnetosphere and it is a source of induced fields driven by variations in the planetary field. A saline ocean would have a sufficiently high electrical conductivity to explain this effect, but the glacial cover would not do so because ice would be almost salt-free and a poor conductor. But, of course, the observations indicate only the presence of a conductor and not its composition.

Satellites are a normal feature of the Solar System, as evidenced by their large numbers for the giant planets. Pluto has a large satellite (Charon), as well as two smaller ones, and the asteroid Ida is seen to have a satellite (Dactyl). We need a special explanation for their fewness in the inner Solar System and this is provided by tidal friction (see Chapter 8, especially Section 8.6, and the comment on the early history of the Moon in Section 1.15).

1.6 Asteroids

The small bodies with orbits concentrated between Mars and Jupiter are sometimes referred to as minor planets, but we prefer to reserve the word planets for the eight large bodies. The word asteroid is the normal scientific term. A few of them have elliptical orbits extending as far as the Earth and are referred to as near Earth asteroids (NEAs) or, sometimes, as the Apollo group of asteroids. They are of particular interest because they are the best observed and because meteorites are NEAs intercepted by the Earth. They may not be totally representative of the larger asteroidal population, and it is possible that some are residual cores of comets. More than 10 000 asteroids have been identified and new discoveries occur at a rate of about one per day. The total number must be very much larger because the population is biased towards small bodies, but only the larger ones are seen. Except for recent collision fragments, the lower size limit is probably set by the Poynting–Robertson and Yarkovsky effects (Section 1.9).

Orbits of the asteroids do not form an uninterrupted continuum but have gaps, known as Kirkwood gaps after their discoverer. The gaps are swept clear of asteroids by resonant gravitational interactions with Jupiter. The 3:1 resonance, for an asteroid with an orbital period 1/3 of the orbital period of Jupiter, has attracted particular attention. A calculation by Wisdom (1983) showed that, for an asteroid in this situation, the Jupiter interaction rapidly increased the eccentricity of the orbit, and Wetherill (1985) argued that this process maintained a flux of fresh asteroidal material in the vicinity of the Earth. The idea is that collisions in the main asteroidal belt project fragments into this and other gaps, so that their orbits evolve until interrupted by gravitational encounters with Mars, the Earth, or perhaps even Venus. They may then be deflected into orbits that evolve more slowly and from which they may be captured by one of these planets.

Collisions in the main asteroidal belt could not be violent enough to project fragments directly into Earth-crossing orbits. The resonant interaction with Jupiter is necessary for maintenance of the population of NEAs against losses by capture, orbital evolution out of range or, in the case of very small bodies, space erosion. However, it is not a sufficient explanation. Bottke *et al.* (2005) pointed out that direct injection of fragments into resonant orbits is too rare to explain the population of NEAs and would produce them only at infrequent intervals. They appealed to the Yarkovsky effect, which brings collision fragments into resonance more slowly, as explained in Section 1.9.

1.7 Meteorites: falls, finds and orbits

Meteorites are iron and stone bodies that arrive on the Earth in small numbers, on elliptical orbits that extend from the main asteroidal belt. Observed falls (firefalls or bolides) are signalled by fiery trails through the atmosphere. A few meteorite falls have been observed in sufficient detail to allow reliable calculations of their pre-terrestrial orbits. This requires timed photographs of the trails from several well separated points. The first clear example was the chondritic meteorite, Pribram, that fell in Czechoslovakia in 1959 and this is one of the five with orbits plotted in Fig. 1.2. Similar orbits are obtained for the larger number of photographed bolides from which there are no recovered meteorites and, more qualitatively, from eyewitness reports of bolides associated with recovered meteorites. An interesting statistical consequence arises from the orbits of meteoritic bodies: falls occur twice as frequently between noon and 6 pm local time as between 6 am and noon, when the opportunity for observation is similar (Wetherill, 1968). This requires the bodies to be overtaking the Earth in orbit when intercepted. It is statistical confirmation, with much larger numbers, of the conclusion from direct observations of bolides that the meteorite bodies are orbiting the Sun in the same sense as the planets, but on elliptical paths extending much farther out than the Earth. This means that when they reach the Earth they have higher orbital velocities. They are asteroidal collision fragments projected into Earth-crossing orbits by the mechanism discussed in Section 1.9.

It is important to distinguish meteorites from meteors, the briefly luminous trails in the upper atmosphere. Most meteors are produced by small particles, called meteoroids, that never get near to the ground. Although a few meteoroids are probably of meteoritic origin, most are small, friable particles of low density, identified as debris from comets. Like comets (Section 1.13), they approach the Earth from all directions. They are not confined to the plane of the Solar System, or to the direction of its rotation, as are the meteoritic bodies.

There are over 1000 specimens of meteorites that were seen to fall, but they are outnumbered by finds, that is bodies that are obviously meteoritic but were not seen to fall. The world collection of finds increased dramatically with the discovery in Antarctica of many thousand meteorites on areas of bare ice. Many of them could have been moved considerable distances by glacial motion of the ice sheet; cosmic ray exposure measurements (Section 1.8) show them to have been on or in the ice for thousands of years. However, the circumstances of the Antarctic finds ensure that they cannot be confused with terrestrial rocks. For this reason they have become important in identifying unusual types of meteorite and in estimating the relative abundances of the different kinds.

There are various classes of meteorite, all composed of stony material and iron, that is, the materials believed to comprise the mantles and cores of the terrestrial planets. Iron meteorites are normally 100% metal, but the stony meteorites commonly contain some iron, in some cases sufficient to classify them as stony-irons. Many stony meteorites contain small rounded inclusions, called chondrules, several millimetres in size, that are chemically distinct from the surrounding material. These meteorites are termed chondrites. Chondrules resemble droplets and, although they probably formed as direct

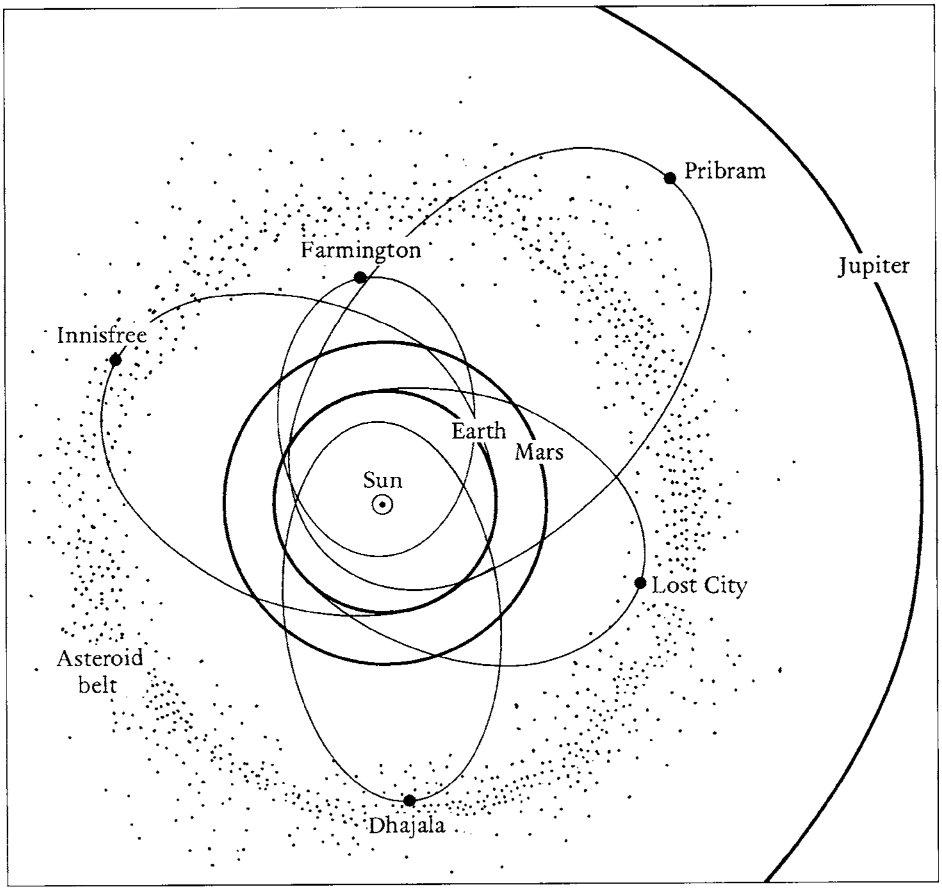


FIGURE 1.2 The calculated orbits of five recovered meteorites identify them with the asteroidal belt. The orbits are drawn to scale, but their orientations are chosen for clarity of illustration. Reproduced by permission from McSween (1999).

condensations of solid from vapours in the solar nebula, they were subjected to subsequent transient heating and even melting. They are uniquely meteoritic, with no equivalent in terrestrial rocks. Chondrites have escaped strong heating and metamorphism that would have converted them to crystal structures similar to terrestrial rocks. Carbonaceous chondrites are a special class, being the meteorite type that is apparently closest to the original accumulation of particles and dust in the solar nebula. As the name implies, they are rich in carbon compounds, which have mostly been lost by the more processed bodies. They also contain refractory inclusions rich in calcium and aluminium (CAIs), that are distinct from chondrules but of similar sizes and appear to have condensed in the solar nebula even earlier than the chondrules. Achondrites are stony meteorites without chondrules that exhibit post-formation metamorphism and differentiation. They have little iron content and are fully crystalline like terrestrial rocks.

The grains and dust in the early solar nebula are believed to have been similar in composition to the carbonaceous chondrites, in which the iron occurs as oxides, especially magnetite, Fe_3O_4. The other meteorite types evolved from this mix. The abundance of carbon allows the suggestion (Section 2.2) that meteoritic iron (and the core material of the terrestrial planets) originated in reactions, similar to that in a blast furnace, triggered by collisional heating. Then the processed material accreted into larger bodies that included metallic iron. The iron meteorites are collision fragments of bodies that had developed sufficiently towards the formation of planets for gravitational separation of iron cores. They have large crystal sizes, evidence of slow cooling and burial in bodies several kilometres in size (Section 1.10).

1.8 Cosmic ray exposures of meteorites and the evidence of asteroidal collisions

Cosmic rays penetrate only the outer 1 m or so of each independent body, so that each asteroidal fragmentation event exposes fresh material to cosmic ray bombardment. Extremely energetic cosmic ray protons cause violent disruption (spallation) of the atomic nuclei in exposed meteorites. A representative example of nuclear spallation is

$$\begin{aligned} {}^{56}\mathrm{Fe} + {}^{1}\mathrm{H} \rightarrow {}^{36}\mathrm{Cl} + {}^{3}\mathrm{H} + 2{}^{4}\mathrm{He} + {}^{3}\mathrm{He} \\ + 3{}^{1}\mathrm{H} + 4\mathrm{n}. \end{aligned} \tag{1.7}$$

Many products arise from numerous similar reactions. When a meteorite arrives on the Earth and is protected by the atmosphere from further exposure, it has accumulated cosmogenic (cosmic ray produced) nuclides from which the duration of its exposure can be determined. Nuclides, with half lives much shorter than the duration of cosmic ray exposure (^{39}Ar, ^{14}C, ^{36}Cl), are maintained in equilibrium concentrations during the exposure but decay after arrival. Their residual concentrations provide a measure of the time that has elapsed since a meteorite arrived on the Earth. This is sometimes referred to as a terrestrial age. Added to the exposure duration it dates the fragmentation event, referred to as the cosmic ray exposure age. Of course these ‘ages’ must not be confused with the age as normally understood, which is the solidification age of original formation, a subject of Chapters 3 and 4.

With correction for terrestrial age, or from measurements on observed falls, uncertainties in cosmic ray intensities and partial shielding of samples by burial in a large meteorite can be allowed for by comparing concentrations of two cosmogenic nuclides, one stable and the other short lived. The concentration of the short lived species is a measure of the rate of production. By selecting pairs of nuclides whose production cross-sections have similar dependences on cosmic ray energy we have two isobaric pairs (^{3}H–^{3}He and ^{36}Cl–^{36}Ar) and two isotopic pairs (^{38}Ar–^{39}Ar and ^{44}K–^{41}K) as species of greatest interest. The first three of these pairs, being gases, also avoid the problem of initial composition that arises in the case of non-volatile spallation products. Then, in terms of the measured concentrations S, R of the stable and radioactive nuclides and their production cross sections σ_S, σ_R determined from laboratory data, the cosmic ray exposure age of a meteorite is given by

$$t = \frac{S}{R}\frac{\sigma_R}{\sigma_S}\frac{t_{1/2}}{\ln 2}, \tag{1.8}$$

where $t_{1/2}$ is the half-life of the active nuclide, which is assumed to be short compared with t. In the cases of the isobaric pairs the stable nuclides are produced by decay of the active ones as well as directly, so that the equation becomes

$$t = \frac{S}{R}\frac{\sigma_R}{\sigma_S + \sigma_R}\frac{t_{1/2}}{\ln 2} \tag{1.9}$$

(Problem 3.2, Appendix J).

Most of the reliable exposure ages for stony meteorites are grouped around 4×10^6 and 23×10^6 years, but others cover the range from 2.8×10^6 to 100×10^6 years with obvious groupings of different types. Some of the lower estimates are probably invalidated by diffusion losses because the same meteorites have small potassium–argon solidification ages. Iron meteorites have generally had much greater exposures, up to a maximum of 2200×10^6 years with groupings at 630×10^6 and 900×10^6 years but not at 23×10^6 years (Anders, 1964). On this time scale variability of the cosmic ray flux can be recognized (Pearce and Russell, 1990), requiring a correction to age estimates. There is a wide scatter and, in some cases, imperfect agreement between different measurements, but there are no coincidences of exposure ages for irons and stones. It is evident that the meteorites are fragments of several, and probably many different bodies, and resulted from impacts that occurred sufficiently long ago to have allowed major modification of the orbits (Section 1.9).

The asteroidal collisions that produced iron meteorites were generally much earlier than those that yielded chondrites, but we must suppose that the more abundant chondritic parents were also involved in early collisions. The most plausible explanation is that orbital evolution of asteroidal impact fragments begins with the the Yarkovsky effect of solar radiation on rotating bodies (Section 1.9). This depends on their thermal conductivities, which are much higher for irons than for stones (see Section 19.6), making orbital evolution slower for the irons. They therefore take longer to reach one of the orbits of gravitational resonance with Jupiter and it is only when they do so that they are projected into Earth-crossing orbits. Thus the grouping of exposure ages alone is not convincing evidence that the different meteorite types came from different parents. However, the wide range of cooling rates apparent from Widmanstätten diffusion zones (Section 1.10), the different chemical histories, as seen in oxidation states, and different ratios of elements make it evident that there were never fewer than several asteroidal bodies. It is probable that there has always been a very large number.

An interesting confirmation that asteroidal collisions project fragments into Earth-crossing orbits is presented by Farley *et al.* (2006), who attributed a pulse of interplanetary dust particles (IDPs) to a fragmentation event 8.3 ± 0.5 million years ago. The event is identified with a 'family' of asteroids with orbits that can be traced back to a common point at that time. They are not close enough to an orbital resonance with Jupiter for any of the major fragments to enter an Earth-crossing orbit, but dust from the initial impact, and following collisions between fragments, spiralled in by the Poynting–Robertson effect (Section 1.9) and some of it was collected by the Earth. Its signature is recognized in marine sediment 8.2 to 6.7 million years old by a peak in the concentration of ^{3}He, which was produced by cosmic ray bombardment of the dust.

1.9 The Poynting–Robertson and Yarkovsky effects

Solar radiation modifies the orbits of small bodies in the Solar System. There are two related effects. Particles that are small enough or are rotating fast enough to remain isothermal, in spite of being irradiated on one side, re-radiate isotropically in all directions. This means that the radiation carries away angular momentum corresponding to the speed of orbital motion. The loss of angular momentum by the particles causes them to spiral towards the Sun. This is the Poynting–Robertson effect, first presented in classical form in 1903 by J. H. Poynting and given a relativistic explanation in 1928 by H. P. Robertson. The complete theory is presented by Lovell (1954), who used it to explain the size-filtering of meteoroid streams, identified as cometary debris. This effect, operating on the fine grains in the early solar nebula, offers an explanation for some of the compositional and isotopic gradients in the Solar System that are otherwise paradoxical (Section 4.5). Larger bodies, rotating slowly enough to be hotter on their sunlit sides and with thermal inertia keeping them hotter on their afternoon hemispheres than on the morning hemispheres that have cooled 'overnight', radiate anisotropically. The modification of their orbits by the radiated angular momentum is the Yarkovsky effect, first discussed about 1900 in an obscure pamphlet by I. O. Yarkovsky. It is comprehensively reviewed in the context of asteroid dynamics by Bottke *et al.* (2005). The essential physical difference between these effects is that the Poynting–Robertson effect depends on $(v/c)^{2}$, where v is the orbital speed and c is the speed of light, but the Yarkovsky effect depends on v/c and can therefore be much stronger, even influencing noticeably the orbits of bodies of kilometre size.

It is convenient to distinguish several effects of solar radiation pressure, although they are not really independent.

(i) There is an outward force from the Sun, opposing the gravitational attraction. A simple interpretation suggests that for particles smaller than a few hundred nanometres the net force would be outwards, blowing them out of the Solar System, but, for particles of sizes comparable to the wavelength of light, the effective optical cross-sections are smaller than the physical cross section and the radiation pressure is reduced.

(ii) A particle on an elliptical orbit receives solar radiation that is Doppler shifted, causing a greater radiation pressure by blue-shifted light on the approaching limb of the orbit than by the red-shifted light on the receding limb. This unbalances the central gravitational force that maintains the particle on an elliptical orbit (Appendix B), and gradually converts the orbit to a circular one.

(iii) For a particle radiating isotropically the orbital angular momentum is progressively transferred to the radiation field because the particle receives radiation with only radial momentum from the Sun but re-radiates it with a forward momentum corresponding to its own motion. The re-radiated light is blue-shifted in the forward direction but red-shifted backwards. The particle loses orbital angular momentum and spirals in towards the Sun. Effects (ii) and (iii) constitute the Poynting–Robertson effect.

(iv) When some of the received radiation is absorbed and re-emitted (at infra-red wavelengths corresponding to surface temperature) gradually during rotation, there is an imbalance of the radiation in the forward and backward orbital directions, because the most recently heated face radiates more strongly. For a body rotating in the same sense as its orbital motion the recoil from the emitted radiation is an accelerating force, causing the orbit to expand. A body with retrograde rotation loses angular momentum and spirals inwards. This is the Yarkovsky effect. Bottke *et al.* (2005) discuss more general cases, including arbitrary orientations of rotation axes, and also the dependence of the effect of albedo, surface emissivity, thermal diffusivity, size and rotation speed.

We examine the principles of these effects with simplified situations. Consider first the Poynting–Robertson effect on a spherical particle of mass m and diameter d in a circular orbit of radius r. Equating centripetal force to the gravitational attraction to the Sun, mass M, the orbital speed is

$$v = (GM/r)^{1/2}, \tag{1.10}$$

where G is the gravitational constant. The total orbital energy, E, is the sum of the gravitational potential energy and kinetic energy and, with substitution for v by Eq. (1.10), is

$$E = -GMm/r + mv^2/2 = -GMm/2r. \tag{1.11}$$

In time dt the particle receives solar radiation energy dε and (by $E = mc^2$) this causes an increase in mass

$$\mathrm{d}m = \mathrm{d}\varepsilon/c^2, \tag{1.12}$$

c being the speed of light. The solar radiation carries no orbital angular momentum, so the angular momentum of the particle is unchanged by this process, that is

$$\mathrm{d}(mvr) = m\mathrm{d}(vr) + vr\mathrm{d}m = 0, \tag{1.13}$$

so that, by Eq. (1.12),

$$m\mathrm{d}(vr) = -vr\mathrm{d}m = -vr\mathrm{d}\varepsilon/c^2. \tag{1.14}$$

The particle re-radiates the energy, dε, isotropically in its own frame of reference, so that there is no reaction on it and its orbital speed is unaffected, but it loses mass dm and therefore angular momentum vrdm. Taking absorption and re-radiation processes together, m is conserved and vr decreases. The rate of loss of angular momentum to the radiation field may be equated to a retarding torque

$$L = m\mathrm{d}(vr)/\mathrm{d}t = -(vr/c^2)\mathrm{d}\varepsilon/\mathrm{d}t \tag{1.15}$$

with a rate of loss of orbital energy

$$\mathrm{d}E/\mathrm{d}t = Lv/r = -(v^2/c^2)\mathrm{d}\varepsilon/\mathrm{d}t. \tag{1.16}$$

The rate at which the particle receives radiation energy is

$$\mathrm{d}\varepsilon/\mathrm{d}t = SA(r_\mathrm{E}/r)^2, \tag{1.17}$$

where $A = (\pi/4)d^2$ is the cross-sectional area of the particle and $S = 1370\,\mathrm{W\,m^{-2}}$ is the solar constant, that is the radiation energy at the radius of the Earth's orbit, r_E. Equating the derivative of Eq. (1.11) and Eq. (1.16) with substitution of Eq. (1.17), we have

$$(GMm/2r^2)(\mathrm{d}r/\mathrm{d}t) = -(v/c)^2 SA(r_E/r)^2 \quad (1.18)$$

and since v is given in terms of r by Eq. (1.10) we have a differential equation for $r(t)$,

$$r\mathrm{d}r/\mathrm{d}t = -2SAr_E^2/mc^2. \quad (1.19)$$

Integrating from the initial condition, $r = r_0$ at $t = 0$, with substitution for A and m in terms of d and the particle density, ρ,

$$(r_0/r_E)^2 - (r/r_E)^2 = (6S/\rho c^2)(t/d). \quad (1.20)$$

The fine grains of interest generally have low densities. Assuming $\rho = 2500\,\mathrm{kg\,m^{-3}}$, expressing t in years and d in mm,

$$(r_0/r_E)^2 - (r/r_E)^2 = 1.16\times 10^{-6}\ t(\text{years})/d(\text{mm}). \quad (1.21)$$

We can apply Eq. (1.21) to the pulse of interplanetary dust grains discussed in Section 1.8. Their starting point was at $r_0/r_E = 3.17$, so that to reach the Earth ($r/r_E = 1$) the numerical value of the left-hand side of this equation is 9.05. Although there is some uncertainty in the time of the asteroidal impact that produced them, from the report by Farley *et al.* (2006) it appears that the transit time to the Earth of the fastest (smallest) particles was no more than 10^5 years, but that the largest ones took about 1.6×10^6 years. Noting that detection of the particles was by cosmic ray-produced ^{3}He, and that the slower particles had more time in space to accumulate it, the 1.6 million year limit can be presumed secure, although the lower limit is less so. Applying Eq. (1.21) to these times, the particle sizes are 0.2 mm to about 0.01 mm. Although many of the particles were probably products of secondary fragmentations and started their independent lives late, this does not invalidate the size estimates. The upper size limit, being more secure, calls for an explanation. It is probably attributable to heating and loss of ^{3}He by grains larger than 0.2 mm during atmospheric entry. Larger grains probably continued to arrive but are not apparent from ^{3}He.

The Poynting–Robertson effect also offers a possible explanation for some of the isotopic variations in the Solar System. Grains surviving from their pre-Solar System histories and incorporated in carbonaceous chondrites have various isotopic ratios reflecting their different nucleo-synthetic sources (Section 4.5). But there is a gradient in oxygen isotopic ratios in the inner Solar System that cannot be explained in this way. If the grains with different origins in the early solar nebula had different size distributions, so that there was a correlation between size and composition, then size filtering by the Poynting–Robertson effect would take effect as a compositional gradient.

When we consider the larger fragments in the asteroidal belt, the Poynting–Robertson effect is too weak to have any influence, and to explain non-gravitational orbit variations we appeal to the Yarkovsky effect. Now we are considering bodies into which heat diffuses on their sunlit sides and is lost from the dark sides. As a surface cools so the heat radiated from it diminishes, so that more heat is radiated from the recently heated 'afternoon' hemispheres than from the 'morning' sides that have cooled 'overnight'. Recoil from the radiation applies a net force to the 'afternoon' hemisphere, causing either acceleration or retardation of the orbital motion according to the direction of rotation. To emphasize the essential physics, we postulate an oversimplified model that exaggerates the magnitude of the effect, but shows why it is fundamentally different from the Poynting–Robertson effect. We ignore the reflectivity/emissivity of the surface, the thermal diffusivity of a body and its rate of rotation and suppose that all of the sunlight falling on it is absorbed and re-radiated after a 90° rotation about an axis normal to the orbital plane. If the body is a sphere of diameter d in an orbit of radius r the rate at which solar energy is received is given by Eq. (1.17) and the rate of momentum transfer to it by the reradiation is

$$F = m\,\mathrm{d}v/\mathrm{d}t = \pm(1/c)\mathrm{d}\varepsilon/\mathrm{d}t. \quad (1.22)$$

This is the radiation force acting on the body, with alternative signs for bodily rotation in the same sense as the orbital motion (+) or the opposite sense (−). The rate of change in

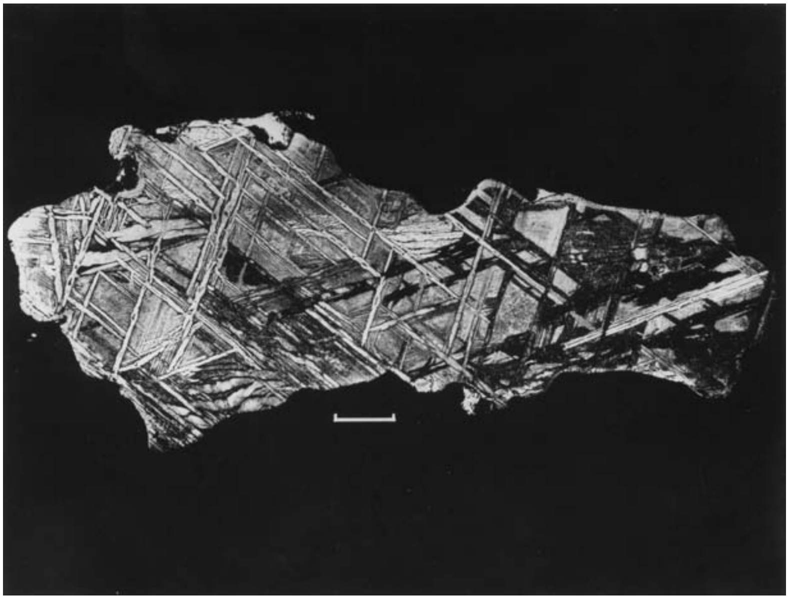


FIGURE 1.3 Widmanstätten structure of the metal phase in the Glorietta Mountain pallasite (stony-iron meteorite). The scale bar is 1 cm. Photograph courtesy of J. F. Lovering.

orbital energy is the product of this force and the speed

$$dE/dt = Fv = \pm(v/c)d\varepsilon/dt. \tag{1.23}$$

Comparing Eqs. (1.23) and (1.16), we see why, in a favourable situation, the Yarkovsky effect can be so much stronger than the Poynting–Robertson effect. For a body orbiting in the asteroidal belt $v/c \approx 6 \times 10^{-5}$.

The words 'in a favourable situation' draw attention to the fact that this simple treatment exaggerates the Yarkovsky effect by assuming that all incident radiation is absorbed and that it is all reradiated at 90°. A body that is rotating rapidly allows no time for establishment of temperature differences around any line of latitude; conversely, a body presenting a constant face to the Sun is subject only to an outward radiation force. Neither experiences a Yarkovsky force. Similarly there is no unbalanced force on a particle that is small enough (and a good enough thermal conductor) to remain isothermal, whether rotating or not. A strong Yarkovsky effect requires particular combinations of body size, rate of rotation and thermal diffusivity, as well as depending on surface properties, albedo and infra-red emissivity. Bottke *et al.* (2005) point out that the effect is strongest when the thermal skin depth for a temperature wave with the rotation frequency $(2\eta/\omega)^{1/2}$ – see Eq. (20.24) – is comparable to a body's dimensions. This means meteorite-sized bodies, but very slow changes affecting bodies of kilometre sizes are also of interest to the evolution of asteroid orbits.

As in Eqs. (1.22) and (1.23), the Yarkovsky effect may have either sign, depending on the rotation of an orbiting body. Its orbital radius may either increase or decrease. Thus it is possible for asteroidal collision fragments to approach resonant orbits from either direction. Bottke *et al.* (2005) argue that this is necessary to the maintenance of a population in highly elliptical orbits (NEAs and meteorites). The slow process of orbital evolution by the Yarkovsky effect precedes the rapid increase in orbit ellipticity by gravitational resonance. Direct injection into resonant orbits is too rare and could produce only occasional NEAs soon after impact events. The argument is reinforced by evidence from the cosmic ray exposures of meteorites (Section 1.8). Their sojourn in space appears to be too long for direct injection. The generally longer cosmic ray exposures of iron meteorites can be attributed to their higher thermal diffusivities, requiring them to be larger for a strong Yarkovsky effect and therefore having orbits that evolve more slowly.

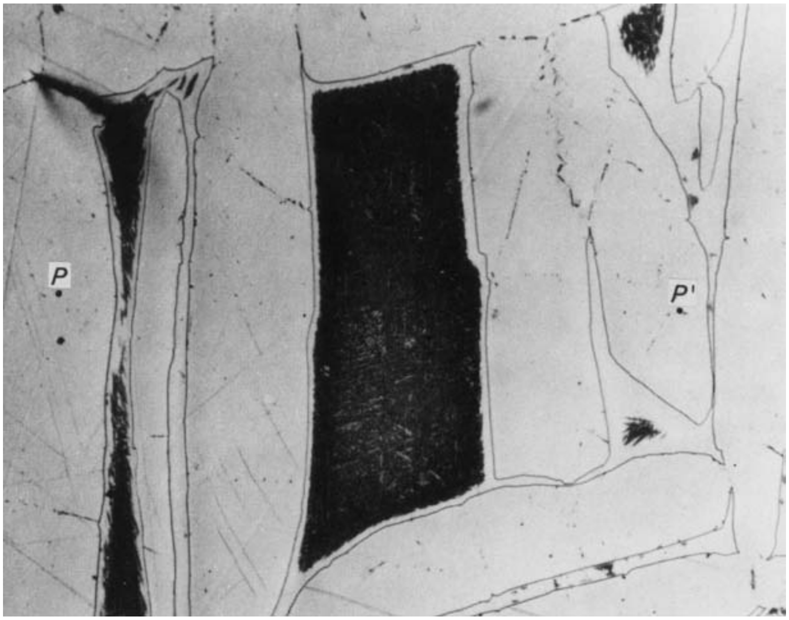


FIGURE 1.4(a) Microscopic picture of a polished and etched slice of the Anoka octahedrite (iron meteorite) showing details of the Widmanstätten pattern. The dark areas are plessite, rimmed by taenite, within the more uniform kamacite.

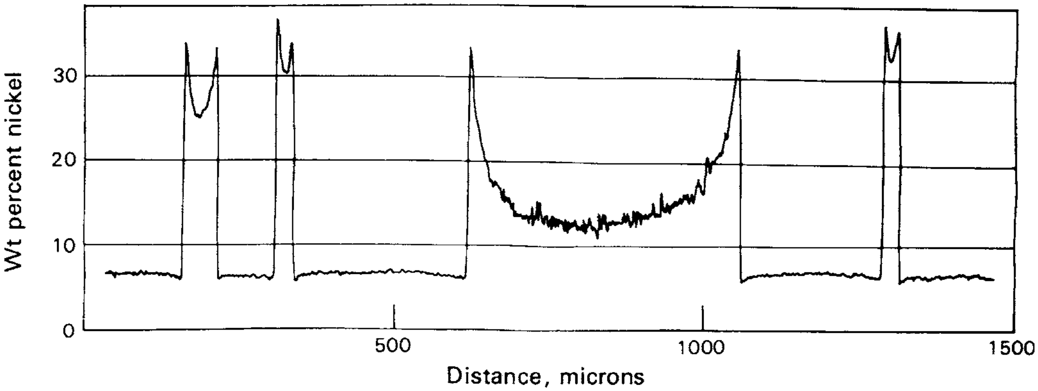


FIGURE 1.4(b) Profile of the variation of nickel content along the line PP′ in Fig. 1.4(a), as determined by an electron microprobe. (This gives chemical analysis of small areas of the surface in terms of the intensities of characteristic X-rays excited by a sharply focussed electron beam.) (Wood, 1964.)

1.10 Parent bodies of meteorites and their cooling rates

Iron meteorites have alloying nickel in solid solution averaging about 9%, but at ambient temperatures there is no single phase with this composition. Two metal phases occur, the body-centred cubic (α) form (kamacite) with 5.5% to 7% Ni, and the face-centred cubic (γ) form (taenite) with variable nickel content, generally exceeding 27%. Both phases occur in close association, as a phase separation from solid solution after solidification of a single phase from the melt. The pattern of interwoven phases is rendered obvious by etching polished sections, as in the example in Fig. 1.3, and is known as the Widmanstätten structure. A third 'phase' – plessite – is also common, but is really a very fine exsolution (phase separation) of kamacite and taenite within the taenite zones, as illustrated in Fig. 1.4. Common crystal orientations across meteorites indicate that the

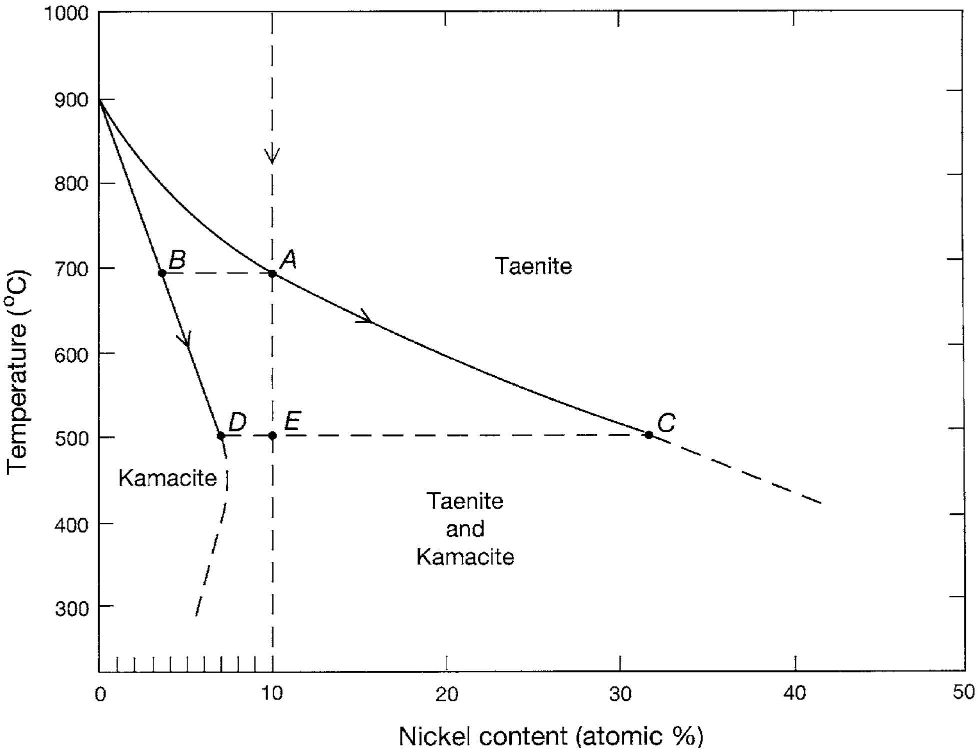


FIGURE 1.5 Phase diagram of nickel–iron according to Goldstein and Ogilvie (1965), with paths followed by exsolving kamacite and taenite in an alloy with 10% Ni, representative of iron meteorites.

original metal crystals were very large, at least one metre across. This indicates slow cooling.

Quantitative evidence of the cooling rate has been obtained from the variations in composition across the kamacite–taenite phase boundaries. The exsolution may be understood in terms of the phase diagram of nickel in iron (Fig. 1.5). An 8% or 10% nickel-in-iron alloy solidifies as taenite and remains as a single-phase solid as it cools down to about 690 °C (point A in Fig. 1.5). At this point it enters the two-phase region, and kamacite, with composition represented by point B, begins to exsolve along {111} planes in the taenite crystal lattice. {111} planes form an octahedral pattern and iron meteorites with well developed Widmanstätten structures are termed octahedrites. The lines AC and BD are phase boundaries and between them is a region of phase instability in which there is no stable single phase. With further cooling to 500 °C the taenite phase increases in nickel content, along path A → C, so forming a decreasing proportion of the alloy, and the planes of kamacite grow in thickness and increase in nickel content, along path B → D. The average nickel content is, of course, still unchanged and is represented by point E, so that at this temperature kamacite has become the major component. With cooling below 450 °C, the solubility of nickel in kamacite begins to decrease. This causes nickel to diffuse out of the boundaries of the kamacite zones into the taenite zones, but diffusion becomes too slow to maintain phase equilibrium and the inhomogeneities in composition shown in Fig. 1.4 develop. Diffusion of nickel is more rapid in kamacite than in taenite, so that the margins of the kamacite are only slightly more depleted in nickel than the core regions, whereas nickel from the kamacite does not penetrate deeply into the taenite. It is the separation of very fine crystals of kamacite in the cores of the taenite zones that forms plessite. This is seen as ‘noise’ on the microprobe record in Fig. 1.4, in which fine details are not resolved.

Widths of Widmanstätten diffusion zones have been used to estimate the rates of cooling of irons and iron-bearing chondrites through the critical range 650 °C to 350 °C. Above this range diffusion is rapid and equilibrium is maintained, and below 350 °C the crystal structure is frozen in. Slow cooling through this range causes wide diffusion boundaries. Estimated cooling rates, from 150 °C to 6000 °C per million years, accord with the large crystal sizes apparent from the consistent orientations of Widmanstätten boundaries over large specimens. Assuming the cooling to have occurred by thermal diffusion in parent asteroids, burial of the samples in bodies a few kilometres or so in radius is implied (Narayan and Goldstein, 1985).

1.11 Magnetism in meteorites

Many meteorites contain fine grains of iron and magnetite, making them suitable objects for paleomagnetic studies (Chapter 25). Iron meteorites do not yield useful results because the large crystal sizes make them magnetically soft. Very fine particles of iron can be magnetically stable, but the most useful observations rely on the presence of magnetite (Fe_3O_4). Intriguing results have been obtained from chondrites, both ordinary and carbonaceous, and from achondrites. Nagata (1979) reviewed the subject and listed a large number of observations. The surprising, but consistent, conclusion is that the chondrites were all exposed to magnetic fields when they were formed. Details of the magnetization processes and the origins of the magnetic fields are still subjects of discussion, but the universality of meteorite magnetizations over-rides the difficulties and doubts about individual cases.

The intensities of the magnetic fields in which the natural remanent magnetizations (NRMs) were induced can be estimated from the NRMs of the meteorites, and their responses to demagnetization by alternating fields and by heat. The usual assumption is that the magnetization is thermoremanent in origin, that is, induced by cooling in a field. The estimates would be changed somewhat by alternative assumptions about the mechanism, but regardless of this uncertainty the general pattern is clear. Ordinary chondrites were exposed to fields between 10^{-6} and 7×10^{-5} T (0.01 to 0.7 Gauss) with a clustering in the range 10^{-5} to 3×10^{-5} T. Except for the irons, their remanences are the least stable of meteorite magnetizations, so there is considerable uncertainty both in the field strength and the inducing mechanism. Although the inference is that they formed in a field, the fact that its strength was comparable to that of the Earth, 3×10^{-5} T to 6×10^{-5} T, invites doubt about the possibility of induction by the Earth's field during or after their arrival. Field intensities for the achondrites were generally rather smaller, but only by a factor two or three. Greatest interest now attends the measurements on carbonaceous chondrites which have greater stability of remanence and were exposed to stronger fields, 10^{-4} T being typical. In this case it is clear that the dominant magnetic carrier is magnetite. Estimates of the inducing field usually assume that the magnetization is thermoremanent in origin, but chemical remanence, resulting from chemical generation or transformation in a field, gives field estimates even greater than the thermoremanence assumption. There is, therefore, no plausible manner in which the remanence of carbonaceous chondrites could have been caused by the Earth's field.

As well as 'whole rock' estimates of the intensities of magnetizing fields, measurements have been made on individual chondrules, especially from the carbonaceous chondrite Allende, which show evidence of two inducing fields. The primary field, which acted on the chondrules individually before their incorporation in larger bodies, may have been as strong as 10^{-3} T, with remanence carried by finely particulate iron (Acton *et al.*, 2007). However, the magnetic moments of the chondrules appear in random directions in the composite material, demonstrating that they were magnetized as independent particles before their incorporation in larger bodies. There is also a secondary chondrule magnetization parallel to that of the whole body. It is less stable than the primary magnetization, allowing the primary field to become apparent after partial demagnetization.

Induction of meteorite magnetizations by the Earth's magnetic field during flight through the atmosphere, or subsequently, must be discounted before alternative explanations can be taken seriously. In this connection it is important that the interior parts of each stone remain cool during atmospheric entry. Although a considerable thickness of material may be ablated away, the remaining heated skin is only a few millimetres thick and is not sampled for magnetic measurements. It is, therefore, not possible that meteorite magnetizations resembling thermoremanence could have been induced by the Earth's field. In any case the Earth's field is not strong enough to have induced the stronger carbonaceous chondritic remanences by any mechanism. We therefore seek an explanation for an extra-terrestrial field or fields, typically 10^{-5} to 10^{-4} T.

As noted in the discussion of Solar System angular momentum in Section 1.4, the Sun is believed to have had a strong and extensive magnetic field during its T-Tauri stage, when the meteorites and planets are assumed to have been forming. It appears possible that the field was strong enough at asteroidal distances to explain meteorite magnetizations, bearing in mind that it would have been intensified by winding up into spirals due to drag of the nebula. We can plausibly argue that the field decreased with time during the formation of meteoritic planetesimals, being strongest during chondrule formation and progressively weaker at the stages when the chondrites and the achondrites were magnetized. But the process of meteorite magnetization by slow cooling or chemical precipitation in such a field faces the difficulty that the meteorite bodies were not stationary with respect to it.

The rotations of the bodies being magnetized, especially those with substantial metal contents, would have been stopped by eddy current damping in the magnetic field, if the field itself had been steady. In any case, rotation in an inducing field merely reduces the effective field to the component along the rotation axis. But we cannot envisage the T-Tauri solar field as conveniently steady and dipolar. The nebular plasma and the field interacting with it could hardly have been less turbulent than the present solar wind and in any case the main solar field could well have reversed as frequently as it does now (approximately once every 11 years). Thus we can reasonably appeal to a solar field of strength adequate to magnetize the meteorites, but not a steady one, and it is necessary to postulate some transient phenomenon as the mechanism for fixing meteorite remanence. T-Tauri-type magnetospheres are known to be subjected to vigorous shock waves. Particularly for the chondrules and carbonaceous chondrites, they are a plausible agent. For ordinary chondrites and achondrites shock compressions due to impacts appear more likely to be effective.

Given the necessity for a transient effect, such as shock compression, the case for an ambient solar field may appear less than compelling, but there are no viable alternatives. Shock hardening of remanence assumes the existence of a field; the shock itself does not provide one. Localized electrical discharges, even lightning, may have been possible if the nebular gas was a sufficiently poor conductor, but lightning does not give rock magnetizations resembling thermoremanence, except where the rock is heated, and so fails to explain the magnetizations of at least many of the meteorites. The suggestion that the parent asteroids had fields generated internally by a mechanism similar to the Earth's dynamo could not account for the random conglomerate-type of magnetizations of the chondrules. Thus, although the explanation of meteorite magnetizations is still tentative, the existence of a magnetic field of order 10^{-4} T in the solar nebula out to asteroidal distances during planetary formation appears impossible to avoid. It justifies the argument in Section 1.4 that angular momentum was imparted to the primordial solar nebula by a magnetic centrifuge mechanism.

1.12 Tektites

Many of the techniques used to study meteorites have been applied to tektites, which are rounded pieces of silica-rich glass, a few centimetres in size, that have been found in tens of thousands over extensive areas of all continents. Potassium–argon and fission track dating (Chapter 3) reveal distinct formation ages (times since fusion) of the different geographic groups. In millions of years the age groupings are: 0.7 (Australia and S.E. Asia), 1.0 (West Africa), 4 (Australia), 14 (Central Europe), 35 (N. America), and possibly 26 (N. Africa). Shapes of the tektites with fused flanges clearly indicate rapid flight through the atmosphere, and the presence of fine metal grains leads to the widely accepted inference that tektites are splashes from massive meteorite impacts. Tektite compositions are consistent with their being the products of impacts on sedimentary rocks. Thus the tektite ages date major meteorite impacts on the Earth for the last 30 million years or so. A lunar origin was hypothesized before lunar samples were available for comparison, but is now discounted.

The strewn fields of tektites extend for thousands of kilometres and there is no possibility that they travelled through the atmosphere for such large distances. The impacts must have been large enough to propel material out of the atmosphere. The very low contents of volatiles, notably water, in the tektites show that they were exposed to a high vacuum before solidification. Their fusion crusts and flanges are products of atmospheric re-entry after solidification in space. However, they fell back rapidly and did not have an extended stay in space, because, although the strewn fields are extensive they are localized and not world-wide. This agrees with the observations of Fleischer *et al.* (1965), who found no cosmic ray-induced fission tracks in tektites and put an upper bound of 300 years on their time outside the atmosphere.

1.13 The Kuiper belt, comets, meteors and interplanetary dust

In 1950 G. Kuiper noted the sharp cut-off in the mass of the known planetary disc at Neptune's orbit (30 AU) and postulated the existence of numerous small objects at greater distances to give a more gradual transition to 'empty' space. Now several hundred objects are known and are appropriately referred to as the Kuiper belt. There is an obvious analogy to the belt of asteroids between Mars and Jupiter. In both cases the distribution of orbits is influenced by gravitational interactions with neighbouring giant planets. As noted in Section 1.6, there are unoccupied gaps in the pattern of asteroidal orbits, the Kirkwood gaps, that correspond to orbital resonances with Jupiter. Similarly, there appear to be several populations of Kuiper belt objects, with a strong clustering in orbits of low eccentricity and semi-major axes in the range 42 to 47 AU, where the Titius–Bode law (Section 1.2) might place a planet. These bodies could well have formed *in situ*. The other major group has very eccentric orbits, with perihelia near to the orbit of Neptune, inviting the inference that they are planetesimals ejected by Neptune from closer orbits, and there is a tight cluster in 3:2 resonance with Neptune (orbiting the Sun twice for every three orbits by Neptune). This third group includes Pluto, which must be seen not as a conventional independent planet but as the first-discovered Kuiper belt object. Its partners in the 3:2 resonant cluster have been dubbed 'plutinos'.

Although the dense population of the Kuiper belt has appeared to cut off at a perihelion distance of about 50 AU, this is not a final cut-off and may simply reflect the difficulty in seeing more distant objects. The discovery in 2003 by Brown *et al.* (2004) of a more distant object, named Sedna, raised new possibilities. Sedna is large by the standards of the Kuiper belt, ~1500 km diameter, 2/3 of the diameter of Pluto, but its closest approach is 76 AU from the Sun and its very eccentric orbit takes it out to about 900 AU, giving it an orbital period exceeding 10 000 years. If there is a population of Sedna-like objects they will be difficult to observe, not only because their remoteness gives them low visibility even at closest approach, but because they spend most time in the even more remote parts of their orbits (Sedna spends less than 1% of the time inside 100 AU). Brown *et al.* (2005) have since reported observations of an even larger body, now named Eris, in a closer but highly inclined orbit. The existence of these bodies invites doubt about the distinctness of the Kuiper belt and whether there is a low density continuum extending much further out, although not as far as the Oort cloud, proposed (also in 1950) by J. Oort as a reservoir of comet nuclei at 75 000 to 150 000 AU. This must be considered distinct because it is not co-rotating with the Solar System. Unlike the planets, asteroids and Kuiper belt objects, comets that wander into the inner Solar System do so from all orientations, as do the trails of debris (meteoroids) that they leave.

Some comets follow elliptical orbits about the Sun, but most approach on orbits that are indistinguishable from parabolas. It is supposed that those now on elliptical orbits were originally following parabolic paths until they suffered gravitational deflection and loss of energy by interacting with the planets, especially Jupiter. The mechanism is the same as the capture of several Jovian satellites. No comets have been observed to 'arrive' on clearly hyperbolic paths, although planetary interactions have caused some to leave with sufficient energy to escape from the Solar System completely. They are commonly considered to be remote components of the Solar System that occasionally stray into the inner parts, but since they do not share the co-rotation of the rest of the Solar System, but approach the Sun from all orientations, apparently randomly and with no preference for the ecliptic plane, the case is less than convincing.

Comets appear to be loose aggregations of millimetre-sized particles frozen into volatile ices that evaporate in the vicinity of the Sun, producing luminous halos and often tails. Without the halos comets are small, very dark and virtually invisible. Ablation of the ices by the Sun releases the small, friable grains that become meteoroids. If they enter the upper atmosphere, they appear as meteors or shooting stars. Meteor showers occur when the Earth passes through bands of orbiting comet debris, but there are also many sporadic meteors that cannot be identified specifically with comets, although most are presumed to be originally of cometary origin. Like comets, they arrive randomly from all directions.

Some information on elemental compositions of meteoroids is obtained from spectroscopic observations of meteors, although their transience restricts the detail and accuracy attainable. Perhaps surprisingly, the compositions are variable and there are characteristic differences between different meteor streams, indicating gross chemical differences between the comets that produced them. The compositions are, overall, compatible with meteoritic abundances, but with more volatiles. The differences are similar to those between different kinds of meteorite. Thus we may suppose the comets to be composed of primitive planetary material, gathered up as dust and volatiles that were driven out of the solar nebula early in its formation.

Another source of primitive planetary material is interplanetary dust particles (IDPs) that are too fine to burn up as meteors in the upper atmosphere ($<20\,\mu$m across), but drift down to the Earth or into the sea. There are two distinct populations of IDPs. In Section 1.8 we refer to the identification of some of them with fragmentation events in the asteroid belt. But others evidently have a more remote origin and appear never to have been incorporated in larger bodies, although they may be related to comets. They are examined individually in dust collected by high flying aircraft. These IDPs are each internally heterogeneous, being composites of even smaller sub-grains with typically meteoritic compositions, silicate, metal, sulphides and carbonaceous material, that preserve the chemical and isotopic signatures of different nucleo-synthetic sources. This was strikingly demonstrated by Mukhopadhyay and Nittler (2004), who reported wide variations in the ^{2}H/^{1}H ratio within a single particle. The very small sizes mean that before capture the particles were strongly influenced by radiation pressure and the Poynting–Robertson effect (Section 1.9). By Eq. (1.25) 20 μm particles of density 2000 kg m^{-3} orbiting in the Kuiper belt at 45 AU would spiral to 1 AU for capture by the Earth in 28 million years (and from there would reach the Sun in another 14 000 years if not evaporated sooner). A simple-minded extrapolation to 4500 million years ago gives a starting distance of 570 AU, inviting the inference that if IDPs are original components of the Solar System they could be dust from a remote part of the Kuiper belt. Implications for chemical and isotopic gradients in the Solar System are discussed in Section 4.5.

1.14 The terrestrial planets: some comparisons

The mean densities of the terrestrial planets and the Moon are listed in Table 1.1. They are sufficiently different to require individual explanations and are not rescaled versions of a standard model. The corrections for self-compression, applied to obtain the estimated zero pressure densities, assume that all of these bodies are composed of iron and silicates, essentially similar to the core and mantle of the Earth, allowing differences in composition, but requiring the materials to follow equations of state for terrestrial materials.

One other simple piece of information that has a direct bearing on the internal structure of a planet is moment of inertia (Problems 1.1, 1.2). This is a measure of the concentration of mass towards the centre, due to both self-compression and the presence of an intrinsically dense core. We still have no information for Venus or Mercury, but, in order of increasing central mass concentration, the known moments of inertia are, with a calculation for the Earth assuming the core and mantle materials to be uniformly mixed together,

Spherical shell	$(2/3)Ma^2$	
Uniform sphere	$(2/5)Ma^2$	
Moon	$0.391Ma^2$	
Mars	$0.366Ma^2$	
Earth	$0.3307Ma^2$	
Homogenized Earth	$0.3727Ma^2$.	(1.24)

The difference between the observed Earth value and that for the same total composition homogenized but still subject to self-compression is a measure of the effect of compositional segregation. It requires a core intrinsically denser than the mantle by a factor of two.

Venus is similar to the Earth in size and mean density and is presumed to be similar internally. The estimated zero pressure density in Table 1.1 is only slightly less than that of the Earth. Since we do not know the moment of inertia and so have no constraint on the internal distribution of density, it is not clear that the difference is real, being due to gross composition, rather than an artifact of the calculation arising from differences in differentiation and depths to temperature-sensitive phase changes. Venus has a core comparable to that of the Earth and it is at least partly liquid (Konopliv and Yoder, 1996), but generates no magnetic field. Its surface is hot (~740 K) and dry. There is a positive correlation between topography and gravity, with no isostatically (hydrostatically) balanced bimodal distribution of surface elevation, corresponding to the continent/ocean basin structure of the Earth. It is probable that Venus has no weak asthenospheric layer, which we can explain by arguing that water is essential to the weakness of the Earth's asthenosphere, and that Venus has no surface water to cycle through an asthenosphere by subduction and volcanism.

Mars is less dense than the Earth and its moment of inertia coefficient, I/Ma^2, is high, indicating that the core is relatively small. However, the total mass and moment of inertia require a mantle of higher intrinsic density than the Earth's mantle. The general reddish colour of Mars and the high iron oxide contents of the SNC achondrites, which are believed to be meteorite impact fragments from Mars (Section 2.5), support the conclusion that Mars has a composition comparable to that of the Earth, but that it is more oxidized and has a smaller total fraction of Fe. Less of the available iron has sunk into the core as metal, leaving the mantle with a high iron oxide content. The magnetic field is weak enough to be explained by remanent magnetism in the thick and highly magnetizable crust, but this required an earlier core-generated field. Connerney *et al.* (1999) reported evidence of magnetic stripes in the Martian crust, apparently similar to those in the ocean floors, implying that Mars once had tectonic activity similar to that on Earth, as well as a self-reversing magnetic field. In that case we can reasonably postulate not only that Mars had an ocean, but that subduction of sea water led to the development of a crust with distinct continental and oceanic areas. Evidence of a bimodal distribution of crustal elevation is now subdued, but may still be there. Yoder *et al.* (2003) reported observations of the deformation of Mars by the solar tide that indicate that its core is still at least partly

Table 1.2 Models of terrestrial planets

Property	Mercury	Venus	Earth	Moon	Mars
r (km)	2440	6051.8	6371.0	1737.5	3389.9
M (10^{24} kg)	0.3302	4.8685	5.9736	0.073 49	0.641 85
$\overline{\rho}$ (kg m^{-3})	5427	5204	5515	3345	3933
I/Mr^2	0.338 ∓ 0.007	0.336[a]	0.3307[b]	0.393_5[b]	0.366[b]
Core density factor	1.05 ± 0.05	1.00	1	1.05 ± 0.05	1.05 ± 0.05
Mantle density factor	0.98 ∓ 0.02	1.00	1	0.971	1.022 ± 0.010
$r_{\text{core}}/r_{\text{planet}}$	0.784 ∓ 0.021	0.522	0.546	0.226 ∓ 0.008	0.422 ∓ 0.017
$M_{\text{core}}/M_{\text{planet}}$	0.679 ∓ 0.015	0.286	0.326	0.024 ∓ 0.002	0.156 ∓ 0.010
Central P (GPa)	38.6 ± 0.5	286	364	5.91 ± 0.06	42.9 ± 1.2
Adiabatically decompressed $\overline{\rho}$ (kg m^{-3})	5017	3868	3955	3269	3697

[a] model calculations
[b] observed

fluid and somewhat larger than the estimate in Table 1.2, but relatively smaller than that of the Earth. Mars has developed massive volcanos, which appear to be supported by the rigidity of a thick, cool lithosphere, but it is not clear that they are currently active.

Mercury has by far the highest uncompressed density of all the terrestrial planets. It is also the smallest, excluding the Moon, so that self-compression is slight, and in particular, the important pressure-induced phase transitions observed in the Earth's mantle would not occur in the depth range of Mercury's mantle. Thus we do not need a value of moment of inertia to deduce that the core of Mercury has about 78% of the planet's radius (Problem 2.2, Appendix J). With respect to oxidation, Mercury is evidently at the opposite end of the scale from Mars. At the modest internal pressures of Mercury we would expect less oxygen dissolved in the metallic core than in the Earth's core. Sulphur may not be a good core candidate either, because of its volatility and the proximity of Mercury to the Sun.

The ratio Fe/(Si + Mg) in Mercury is clearly higher than for the other terrestrial planets and demands fractionation in the early solar nebula. The resulting large core leaves the mantle only 500 km deep. The surface appears to have been disturbed rather little by tectonic processes since massive impact cratering, which we suggest was caused by the infall of fragments from one or more vestigial satellites (Section 1.15). The magnetic field of Mercury is of particular interest. Although much weaker than the fields of the Earth or giant planets, it is more than 100 times stronger than the fields of Venus, Mars or the Moon (Table 24.2). Explanations in terms of crustal magnetization and induction by the solar wind have been considered but appear inadequate and in Section 24.8 we conclude that it is almost certainly driven by an internal dynamo. This requires a core that is at least partly fluid, and Margot *et al.* (2007) reported mechanical evidence that this is so.

The Moon is less dense than the terrestrial planets and it is small enough to have only a 2.3% difference between compressed and uncompressed densities. The central pressure is less than required for the silicate phase transitions of the Earth's mantle, and the observed moment of inertia cannot be explained by self-compression, even though it appears to be close to the value for a uniform sphere (0.4). A compositional variation is required and the obvious explanation is a small metallic core. A core with 2.5% of the mass of the Moon and 22% of its radius

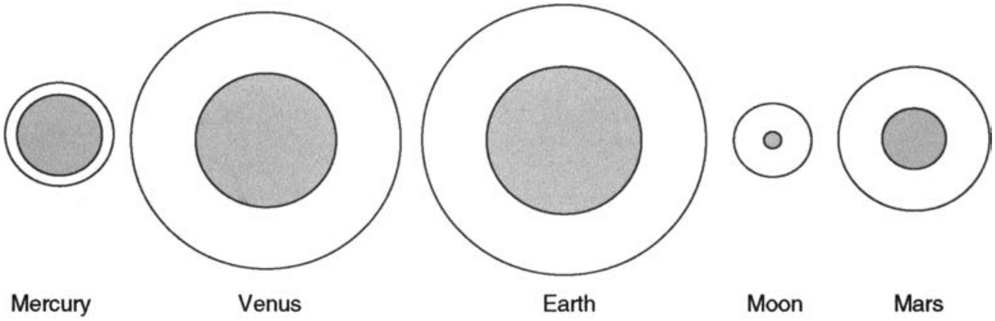


FIGURE 1.6 Relative core sizes of the terrestrial planets and the Moon.

suffices. A heterogeneous mantle or very thick crust cannot be ruled out as alternatives and records from the lunar seismometers did not establish the presence or absence of a core, but the evidence of an early lunar magnetic field demands one. Stevenson (2003) and Williams *et al.* (2004) reported evidence that the core is still at least partly fluid. Thus we conclude that all of the terrestrial planets and the Moon have cores that are partly liquid.

Table 1.2 lists mechanical details of models of terrestrial planets and Fig. 1.6 illustrates the core sizes. These models are results of calculations that assume that the other planets and the Moon have metallic cores and silicate mantles following the same equations of state as the Earth's core and mantle, but with densities that may be slightly different (Stacey, 2005). For Mercury, Mars and the Moon the possibility is allowed that the cores have densities between 1.0 and 1.1 times the density of the Earth's core. This is the meaning of 1.05 ± 0.05 and not that this range represents an uncertainty in the conventional sense. Then with (ρ/ρ_0) vs P matching the equations of state for Earth materials, the core sizes and mantle density factors were adjusted to give the observed planetary radii and masses (and also moments of inertia in the cases of Mars and the Moon). The $\pm$ or $\mp$ ranges correspond to the assumed core density ranges. Thus for Mars the estimated core and mantle density factors are positively correlated but in the case of Mercury the correlation is negative. For Mars and the Moon the observed moments of inertia provide an additional constraint, making the calculations more secure than for Venus or Mercury.

1.15 Early history of the Moon

The origin of the Moon has been a fertile source of conjecture. Even in modern times several quite different hypotheses, including capture from an independent solar orbit and fission of the Earth, have had strong advocates. A fashionable variant of the fission hypothesis is that the Moon accreted from debris thrown up by a massive (Mars-sized) impact on the Earth. The original logic of this idea was essentially geochemical; an inference that the Moon is composed primarily of terrestrial mantle material already segregated from the core. In spite of serious geochemical objections (Ringwood, 1989; Lee *et al.*, 1997) and calculations indicating that it is only marginally possible mechanically (Canup and Asphaug, 2001), this hypothesis has almost become the conventional wisdom, so we review some of the difficulties that it faces. As a general observation, we regard satellites as a normal accompaniment to planets and seek an explanation for their fewness in the inner Solar System, and not a special explanation for the fact that the Earth has one.

There is a gradient in the oxygen isotope ratio, $^{18}O/^{16}O$, in the Solar System, as discussed in Section 4.5 and illustrated in Fig. 4.2. For the Moon this ratio falls on the terrestrial fractionation line, demonstrating that the Earth and Moon formed at the same distance from the

Sun. By the giant impact hypothesis, the lunar composition would be dominated by material from the impactor, which must, therefore have been in a solar orbit close to that of the Earth. This means that its relative velocity was modest and debris from the impact would have formed a moon in a close terrestrial orbit. This agrees with the calculations of Canup and Asphaug (2001), which suggest that the Moon formed just outside the Roche limit of gravitational instability at 3 R_E (Earth radii) (Section 8.5). It would not have formed at about 25 R_E, as required by extrapolation of tidal friction observations (Section 8.6).

We take a critical look also at the commonly advanced argument that the angular momentum of the Earth–Moon system requires a special explanation, such as a giant impact. The orbital angular momentum per unit mass of satellites is necessarily very much greater than the rotational angular momentum of their parent planets. For the Moon the ratio is 400. It is much larger still for the Jupiter system and, in Section 1.4, we point out that for the Solar System as a whole 99.5% of the angular momentum is attributed to orbital motion of the planets, which have about 0.1% of the mass. It is misleading to refer to a high angular momentum of the Earth–Moon system. All that this means is that the Moon has a substantial fraction of the total mass, so that its orbital angular momentum makes a large contribution to the total. If the angular momentum argument is to be pursued it must be diverted to a claim that the large Moon/(Earth + Moon) mass ratio (1.2%) requires a special explanation. Is this ratio anomalous? We believe not. In the Solar System there are rather few examples to use for comparison. Mercury and Venus have lost any satellites they once had, for a reason discussed below. Pluto's satellite, Charon, appears to have about 12% of the mass of that system. Satellites of the giant planets are too small to hold the lighter gases and we cannot make an effective comparison with the heavy elements in their parent planets. That leaves Mars as the anomaly, with so little mass in its satellites. So, if we argue about angular momentum, what data do we have to go on? Perhaps we should consider the rotational speed that a parent body would have if merged with its satellites, conserving angular momentum. For the Earth, that would mean a rotational period of 4 hours. Repeating the calculation for the Solar System as a whole, the rotational period of the Sun would be about 3.5 hours (compared with the present 28 days). We do not suggest that this is a very significant statistic, but only that the angular momentum of the Earth–Moon system offers no evidence for a giant impact.

All of the giant planets have numerous satellites (Table 1.1). Pluto has three (Weaver *et al.*, 2006) and the asteroid Ida has one, but the terrestrial planets have only three between them. Mars has two very small ones and Venus and Mercury have none, making the Earth's single, large satellite exceptional among the terrestrial planets. But there is a straightforward explanation for this situation. The orbit of the Moon is evolving due to the energy dissipated by the tide that it raises in the Earth (Section 8.4). The Earth is losing rotational energy but a small fraction is imparted to the lunar orbit, causing the Moon to recede from the Earth at a rate that has caused a major change in the orbit over its lifetime. The recession of the Moon is a consequence of the fact that the Earth's axial rotation is faster than the Moon's orbital motion. If the Earth were rotating more slowly than the orbital motion, then tidal friction would, instead, cause the Moon to spiral in towards the Earth because the effect of tidal friction is to oppose the relative rotation. The Sun also raises a tide in the Earth, but it is smaller than the lunar tide. However, Venus and Mercury, being closer to the Sun, have larger solar tides. Solar tidal energy dissipation (which varies with distance r from the Sun as r^{-6}) has effectively stopped their rotations. Thus tidal friction would have caused any satellites that they once had to have spiralled in towards them, eventually reaching the Roche limit of gravitational instability (Section 8.5), although the final stage of accretion of their fragments by the parent planets is not so clear. If the Earth's rotation had been stopped by the solar tide, then, from its primordial distance, the Moon would have plunged into the Earth long ago.

Tidal friction also provides an explanation for the fact that the Earth now has just one, large satellite. At the present rate of rotation of the Earth a synchronous orbit would be at 6.64 Earth radii and any satellite outside this range would recede from the Earth. For the faster original rotation of the Earth the synchronous orbit would probably have been inside the Roche limit at 3 Earth radii (Section 8.5), so all of the original satellites would have been receding from the Earth. The rate of recession would have been much faster for the inner ones unless they were very small, varying as $mr^{-5.5}$ for mass m at distance r (Eq. 8.32). We can therefore see that any number of satellites that were not initially too remote would have coalesced into a single body. There is no reason to consider the Earth–Moon system to be anomalous or to require a special event or process. Perhaps we should consider Mars to be anomalous because there is so little mass in its satellites, but the single large satellite of the Earth and the absence of satellites of Mercury and Venus are inevitable consequences of tidal friction.

The cratered lunar surface has preserved a record of its bombardment and the discovery of impact melts in the lunar samples returned by the Apollo missions allowed the impact dates to be estimated. A major, widespread isotopic disturbance 3.9×10^9 years ago was promptly recognized (Tera *et al.*, 1974), leading to the inference of a lunar cataclysm at that time. Identification of the event with an intense bombardment of limited duration has received strong support (Ryder, 1990; Dalrymple and Ryder, 1993, 1996; Ryder and Mojzsis, 1998; Cohen *et al.*, 2000), but also doubts (Hartman, 2003). There are two sources of evidence: impact melts from samples collected by astronauts and from meteorites identified as fragments from later impacts on the Moon. Evidence for impacts before 3.9×10^9 years ago is very limited. Both kinds of sample include melts indicative of dates up to 1.5×10^9 years younger, but this is less common for the Apollo samples. The alternative to the cataclysm hypothesis is that bombardment only gradually declined after formation of the Moon, with no cataclysmic peak, but that earlier evidence was erased by the intense resurfacing. However, this is incompatible with the existence of lunar basalts with ages up to 4.2×10^9 years. If we need to choose between the two data sets we favour the Apollo samples, because lunar meteorites are themselves products of later impacts sufficiently violent to cause some resetting of isotopic clocks by shock wave heating. An objection to our focus on the Apollo samples, for which evidence of the cataclysm may be a little clearer, is that they may represent only localized areas of the lunar surface. But the massive impacts, apparent from the sizes of the lunar craters, would have distributed debris all over the Moon and not just locally. We take the view that evidence for the cataclysm, a massive bombardment of the Moon after 600 million years of relative quiescence, is well documented and demands an explanation.

At this point we depart from the supposition (e.g. Kring and Cohen, 2002) that the same cataclysmic bombardment affected the whole of the inner Solar System. Indeed, if it had done so, the Earth, with stronger gravity, would have been more intensely bombarded, although lack of evidence for this does not carry great weight because very little crust has survived from that time. We consider it far more plausible that the debris that impacted the Moon had been in terrestrial orbit, as a second moon, for 600 million years. With the lunar orbit evolving by tidal friction, as Wetherill (1981) pointed out, the smaller body eventually came within the Roche limit of the larger one (Section 8.5) and broke up; its fragments remained in terrestrial orbits and bombarded the Moon over the next several million years, probably after multiple secondary fragmentations. The analysis in Section 8.6 indicates that this occurred when the Moon was at a distance of about 40 Earth radii (compared with the present 60.3 Earth radii) and that the delay between formation of the Moon and the cataclysm agrees well with the time scale of orbital evolution by tidal friction.

Our variant of the lunar cataclysm hypothesis has some further implications. Since the postulated second moon and its fragments were in terrestrial orbits, not dramatically different from that of the major moon, the impact velocities were modest. The fragments were large and caused massive craters, but they were not moving at speeds comparable to those of asteroidal bodies on highly elliptical orbits. Thus, only a modest fraction of the material ejected by the impacts would have escaped the lunar gravity and most of that would have remained in terrestrial orbit for subsequent collection by the Moon. Very little, if any, would have reached the Earth. This adds further emphasis to our point that the cataclysm affected only the Moon. No material from elsewhere was involved and there was no bombardment of the Earth.

We suppose that late bombardments affected Mercury and Venus, as well as the Moon, but they were independent events. If, as we suggest above, Venus and Mercury once had satellites which were caused to plunge into them by tidal friction, as soon as solar tidal friction slowed the planetary rotation, then the final stage would have involved break-up inside the Roche limit. Each planet would have been impacted by numerous fragments. This is consistent with the cratered surface of Mercury. Late bombardments of this kind require a mechanism that causes orbital evolution and tidal friction is the obvious candidate. As we point out in Section 8.6, it is well understood, although its significance to the history of the inner Solar System has not been fully recognized.

2

Composition of the Earth

2.1 Preamble

The elements of the Solar System are products of several nucleo-synthetic events but were almost completely mixed in the solar nebula before planetary accretion began. Fine grains in carbonaceous chondrites have preserved a record of early events; the final one was a supernova that preceded planetesimal accretion by no more than a million years. Elements heavier than iron and all or most of the radioactive species were supernova products. The non-volatile elements accreted in the inner Solar System, forming the terrestrial planets and the meteorite parent bodies. The mixture is dominated by elements with atomic masses that are multiples of 4, a nuclear structure favoured by the strong binding of ^{4}He nuclei (Table 2.1). This selection of available nuclides places a restriction on hypotheses concerning planetary composition. It adds confidence to our understanding that the same major elements formed all of the terrestrial planets as well as the meteorites, and that the meteorites give us a broad picture of planetary chemistry.

As in the meteorites, the most abundant elements in the Earth are oxygen, iron, magnesium and silicon (masses 16, 56, 24, 32). For the compositional model summarized by Table 2.2, the proportions by mass are 31.5%, 30.3%, 15.4% and 14.2%, with all the other elements together making up the remaining 8.6%. The uncertainties in these numbers are indicated by the differences between proportions of major refractory elements in the meteorite and Earth columns of Table 2.1. The abundance of metallic iron in meteorites allowed it to be recognized as the major constituent of the Earth's dense core as soon as the core was identified by seismology. This accounts for most of the iron in the inventory of the Earth's constituents, leaving an overlying mantle dominated by MgO and SiO_2, with lesser amounts of FeO, CaO, Al_2O_3, Na_2O and others. These oxides form various compounds that comprise the mantle minerals, with pressure controlling the mineral structure. With respect to these major constituents the mantle is believed to be essentially homogeneous but with a physical layering caused by pressure-induced phase transitions to progressively denser mineral structures with increasing depth.

We have direct access to a very small fraction of the Earth, the uppermost part of the crust, and the whole crust constitutes only 0.5% of the mass of the Earth and is not representative of the total Earth composition. The underlying mantle makes up 67% and the core 32.5% of the total. They are composed of materials that are denser than the crust, even allowing for compression. We have, therefore, only indirect methods of determining the overall composition of the Earth. Seismological modelling (Chapter 17), high pressure experiments on candidate materials (Section 18.2), analyses of mantle-derived igneous rocks, comparisons of their radioactive contents with the geothermal flux (Chapter 20) and the existence and behaviour of the geomagnetic field (Chapter 24) give important clues. But our confidence that we are correctly assessing the composition of the Earth really derives

Table 2.1 Estimated relative abundances by mass of elements in the solar photosphere, meteorites and the Earth. Values are normalized to the abundance of silicon. See Newsom (1995) and McDonough and Sun (1995)

Mass of most abundant isotope	Element	Sun	Meteorites	Earth
1	H	1003		
4	He	392	–	–
16	O	13.6	2.2	2.22
12	C	4.4	0.33	
20	Ne	3.5		
56	Fe	2.6	1.81	2.14
14	N	1.6	0.001	20 ppm
28	Si	1	1	1
24	Mg	0.91	0.91	1.09
32	S	0.52	0.60	0.20
36	Ar	0.13	20 ppb	20 ppb
58	Ni	0.105	0.105	0.16
40	Ca	0.092	0.088	0.12
27	Al	0.080	0.082	0.11
23	Na	0.049	0.047	0.013

Table 2.2 The most abundant elements in the Earth: percentages by mass for the mantle and core models in Sections 2.7 and 2.8 with some values for the upper crust, as considered in Section 2.9

Element	Upper cont. crust	Mantle	Outer core	Inner core	Earth
O	46.8	44.23	5.34	0.11	31.47
Fe	3.5	6.26	79.15	84.43	30.26
Mg	1.3	22.80	–	–	15.36
Si	30.8	21.00	–	–	14.15
Ni	–	0.2	6.49	6.92	2.27
S	3.0	0.03	8.84	8.02	2.78
Ca	–	2.53	–	–	1.70
Al	8.0	2.35	–	–	1.58
Na	2.9	0.27	–	–	0.18
Others	3.7	0.33	0.58	0.52	0.22
Mean at. wt.	20.8	21.08	44.53	50.16	25.72

from its similarity to the non-volatile constituents of meteorites and the solar atmosphere. Although the compositions of all the terrestrial planets can be explained by the same basic building blocks, the proportions are not all identical. Some fractionation of elements occurred. A very noticeable example is seen in the density of Mercury, which requires a higher proportion of iron than is possible for the Earth.

There are mineralogical phase transitions in the mantle at 410 km and 660 km, with a less striking one at 520 km. By convention the lower mantle starts at the 660 km boundary and extends to the core–mantle boundary (depth 2890 km). The region above it, referred to as the upper mantle, includes the phase transition zone and is more obviously heterogeneous than the deeper part. The possibility that the upper and lower regions of the mantle are chemically distinct has often been canvassed, but for a variety of reasons the difference is believed to be slight. A debate that has hung on this is whole mantle convection vs separate upper and lower mantle circulations. A suggestion that the 660 km transition inhibits convection through that depth is addressed in the discussion of thermodynamics of mantle convection (Chapter 22). Chemical isolation of upper and lower mantles would introduce considerable compositional uncertainty. Mineral inclusions in diamonds that have evident lower mantle origins and 'hot spot' basalts, believed to be partial melts from convective plumes originating at the base of the mantle (Chapter 12), provide samples consistent with a more or less homogeneous mantle composition. Moreover, seismological imaging indicates that cool lithospheric slabs, subducted from the surface, penetrate deep into the lower mantle. Evidence that the sources of some mantle-derived rocks have maintained chemical (and isotopic) isolation for the entire life of the Earth must be explained without postulating isolated or compositionally distinct upper and lower mantles.

A model of mantle composition consistent with its origin as primitive (type 1 carbonaceous) meteorites, from which volatiles and core constituents have been removed, is referred to as the pyrolite model. It was originally derived by A. E. Ringwood from elemental fractionations in the process of partial melting that produces basaltic magma. Although suggestive of fire, the word pyrolite is a contraction of the two principal minerals, PYRoxene and OLivine, a simple combination of which approximates the mantle composition. While there are numerous variants of the pyrolite model, we can take the low pressure form to be 60% olivine ($(MgFe)_2SiO_4$), 30% pyroxene ($(MgFe)SiO_3$) and 10% garnet ($(FeMgCa)_3Al_2Si_3O_{12}$). The garnet is more close-packed than olivine or pyroxene and so survives compression better. It tends to absorb the others with increasing pressure until more dramatic phase changes convert the minerals to new structures. A detailed development of the pyrolite model of the mantle composition, presented by McDonough and Sun (1995), appears reasonably secure, but there is more uncertainty about the core.

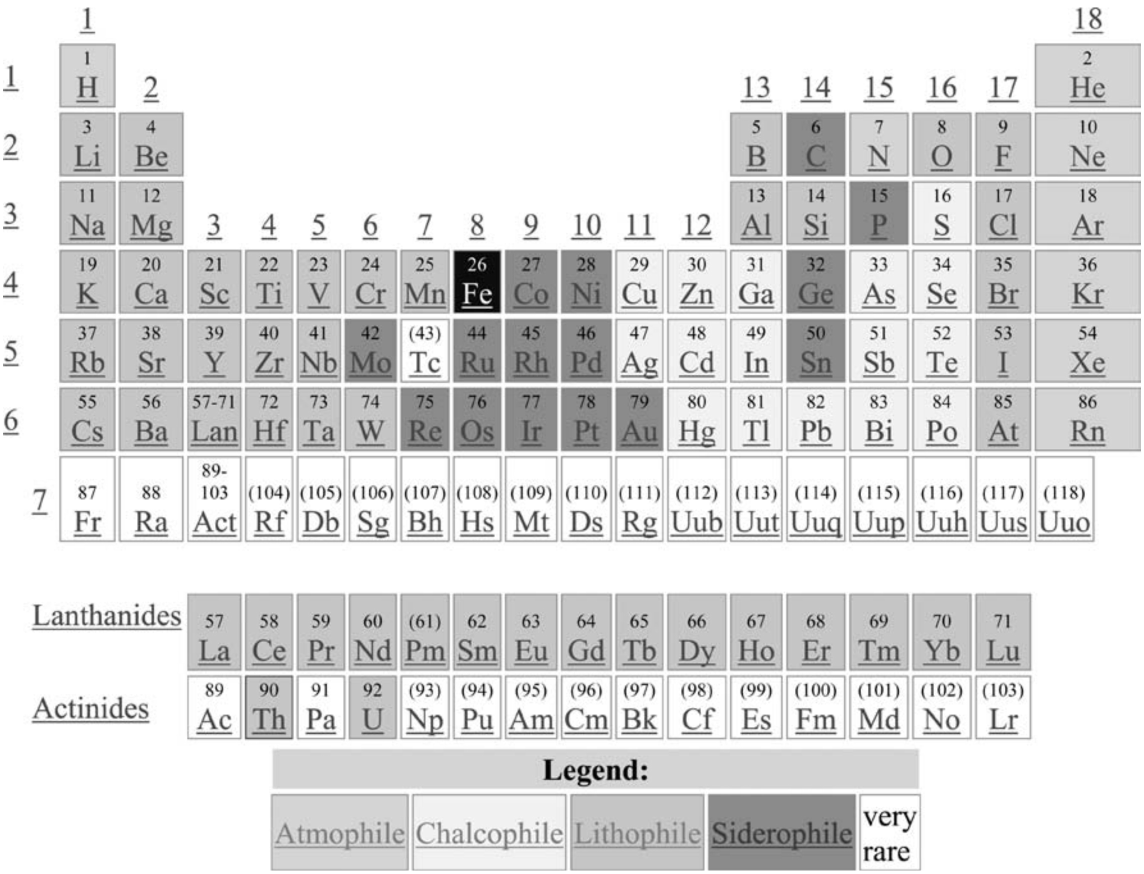
The mantle is not representative of the Earth as a whole because some elements, especially iron, have settled into the core. Although iron is the dominant element, the core is 10% less dense than pure iron under similar conditions and the mixture of lighter elements causing this has been debated for several decades (Poirier, 1994). The serious candidates are, in order of increasing atomic weight and with mass fractions in the outer core required if each were the only light ingredient, H (1.4%), C (10.6%), O (12.7%), Si (17.7%), S (18.2%). The choice between them affects the estimated overall composition of the Earth. Arguments in Section 2.8 favour a mixture primarily of S and O in the outer core, with S but little O in the inner core. The presence of both H and C must be allowed but Si is not favoured. Also we consider that the core is likely to contain more Ni than is suggested by compositions of carbonaceous chondrites and is better represented by the Ni contents of iron meteorites.

We use the mantle + crust (silicate Earth) composition by McDonough and Sun (1995) and an estimated core composition based on Table 2.5 to obtain the resulting total bulk Earth composition in Table 2.2. Also listed is the composition of the upper crust in continental areas, as estimated by McLennan (1995). The crustal composition is very diverse, and in referring to it we emphasize only that its overall average differs from that of the mantle. With respect to the major constituents of the Earth, the contribution by the crust is lost in the uncertainties, but many of the minor elements are concentrated in the crust. Notable are the thermally important radioactive elements, K, U and Th (see Chapter 21). The crust–mantle boundary (Mohorovičić discontinuity) is identified by its density and seismic velocity contrasts. It marks a world-wide compositional difference. The biggest difference between crust and mantle compositions is in the Mg concentration, partly compensated by Al, leaving the crust much richer in Si. In view of the prominence of these elements, the crustal composition is sometimes referred to as sial (Si-Al), to distinguish it from the mantle sima (Si-Mg).

A useful summary of the migration of elements in the evolution of the Earth is their grouping in the periodic table according to geochemical behaviour (Table 2.3). Siderophile (iron-loving) elements that are presumed to be core constituents are tightly clustered in the table, and atmospheric elements are also an obviously distinct category. Lithophile (silicate-loving) elements are left after extraction of the siderophile and chalcophile (sulphur-loving) elements. There is a distinct group of elements, termed ‘incompatible’, that do not fit well into mantle crystal structures and separate into the fluid during partial melting. Volcanic processes concentrate them in the crust. They include all the thermally important radioactive species and leave elements such as Mg in a mantle residuum.

Hydrogen, primarily in the form of water, has several crucial roles in the Earth, although it is not represented with the most abundant elements in Table 2.2. It is especially obvious at the surface, 70% of which is covered by water, and occurs in trace abundance throughout the mantle. Water is cycled through the atmosphere at a rate equivalent to the volume of the oceans in about 3000 years. Also, ocean water is cycled through the uppermost mantle, being carried

Table 2.3 Classification of elements by geochemical behaviour according to V. Goldschmidt



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	1 H																	2 He
2	3 Li	4 Be											5 B	6 C	7 N	8 O	9 F	10 Ne
3	11 Na	12 Mg											13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
4	19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
5	37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	(43) Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
6	55 Cs	56 Ba	57-71 Lan	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
7	87 Fr	88 Ra	89-103 Act	(104) Rf	(105) Db	(106) Sg	(107) Bh	(108) Hs	(109) Mt	(110) Ds	(111) Rg	(112) Uub	(113) Uut	(114) Uuq	(115) Uup	(116) Uuh	(117) Uus	(118) Uuo

Lanthanides	57 La	58 Ce	59 Pr	60 Nd	(61) Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb	71 Lu
Actinides	89 Ac	90 Th	91 Pa	92 U	(93) Np	(94) Pu	(95) Am	(96) Cm	(97) Bk	(98) Cf	(99) Es	(100) Fm	(101) Md	(102) No	(103) Lr

Legend:				
Atmophile	Chalcophile	Lithophile	Siderophile	very rare

down in subduction zones with sea-floor sediment. Tectonics as we know it depends on this cycle. Water has a dramatic effect on the mechanical properties of mantle rocks. Although we have only indirect evidence of the abundance of water deep in the mantle from ocean island basalts, the viscosity is not explicable without some water. The occurrence of hydrogen in the core is considered in Section 2.8, but it is not a major component in the compositional model in Table 2.5.

In a discussion of the composition of the Earth we need to note the role of the atmosphere, which is at least partly an outgassing product, and the evidence of its evolution is a record of the history of the Earth as a whole. The Earth is environmentally unique, as is illustrated by a comparison of the atmospheres of Venus, Earth and Mars (Tables 2.8 and 2.9).

2.2 Meteorites as indicators of planetary compositions

As we mention in Chapter 1, meteorites are of several kinds, with compositions ranging from 100% metal to 100% stone, but they are commonly a mixture. Their relevance to the composition of the Earth hinges on evidence that they were derived from a common source in the solar nebula by a variety of events and processes, and that their total composition is similar to that of the nebula from which the terrestrial planets accreted. We consider four observations which, taken together, provide convincing evidence that this is so and invite the obvious conclusion that the Earth and other terrestrial planets formed from the same primordial brew.

(i) Almost all meteorites have a common formation age, about 4.57×10^9 years (Section 4.3).
(ii) With a few interesting but very special exceptions, discussed in Sections 4.5 and 4.6, the meteorites have the same isotopic ratios as one another (and the Earth). Distant molecular clouds, observed spectroscopically, have quite different ratios, being derived from different nucleo-synthetic events.
(iii) Abundances of elements in the solar atmosphere (Table 2.1) are consistent with the overall composition of meteorites (Section 2.6).
(iv) All of the meteorite types can, in principle, be produced from the most primitive (least processed) type, the carbonaceous chondrites (Section 2.4), by heating and reduction with some segregation of the processed material. The densities of the terrestrial planets can also be explained in this way.

2.3 Irons and stony-irons

Many of the meteorites are composed entirely or mainly of metallic iron, alloyed with nickel, and many of the stony meteorites, which are more abundant than the irons, contain some metal. Iron meteorites generally survive hypersonic flight through the atmosphere better than stones and they are more obviously different from common crustal rocks than are stony meteorites. Being more immediately recognizable as extra-terrestrial objects, they are the type meteorites of popular literature. A characteristic feature of irons and stony-irons is an exsolution of two phases of the FeNi alloy, rendered visible by etching polished sections, as in Fig. 1.3. This is the Widmanstätten structure, which provides a measure of the cooling rates, showing that the asteroidal parent bodies of these meteorites were several kilometres in radius (Section 1.10).

It is easy to visualize the iron meteorites as forming the cores of their parent bodies, although this is conjectural, because there are no coincidences in cosmic ray exposure age for irons and stones (Section 1.8). The gravitational separation would have been quite sluggish in a body probably less than 1/1000 of the Earth's radius with correspondingly weaker gravity, and it is not surprising that many of the meteorites are mixed iron and stone with incomplete separation. Unlike the situation in the Earth, the gravitational energy of core separation was a negligible contribution to the heat required for melting. Neither is there convincing evidence for sufficient early short-lived radioactivity and the only obvious sources of heat are the kinetic energy of collisions between merging bodies and chemical energy released by carbon reduction of magnetite to metallic iron. We can understand that the heating and resulting metamorphic processing of meteorite parent bodies were very variable. A plausible mechanism for the production of meteoritic iron would start with a collision between asteroids of primitive carbonaceous material (Section 2.4), containing iron as magnetite (Fe_3O_4) and carbon compounds, as well as silicate. Such a collision would generate sufficient local heat to trigger a blast-furnace reaction, producing metallic iron and blowing off carbon monoxide.

Pallasites are an interesting type of stony-iron meteorite and examples are illustrated in Fig. 2.1. The material with a metallic appearance is troilite (FeS). Sulphur dramatically lowers the melting point of iron. The eutectic temperature of an Fe-S mix is 1261 K, more than 500 K lower than the melting point of pure iron, so that, in the presence of sulphur, melting would occur with heat inadequate to melt Fe or Fe-Ni alloy. This is persuasive evidence that S has accompanied Fe into the Earth's core (Section 2.8). The stony parts of such stony-irons have achondritic mineralogy (Section 2.5).

2.4 Ordinary and carbonaceous chondrites

There are several kinds of chondrite, but collectively they are more abundant than all other types of meteorite taken together. They are stony meteorites, most of which have structures

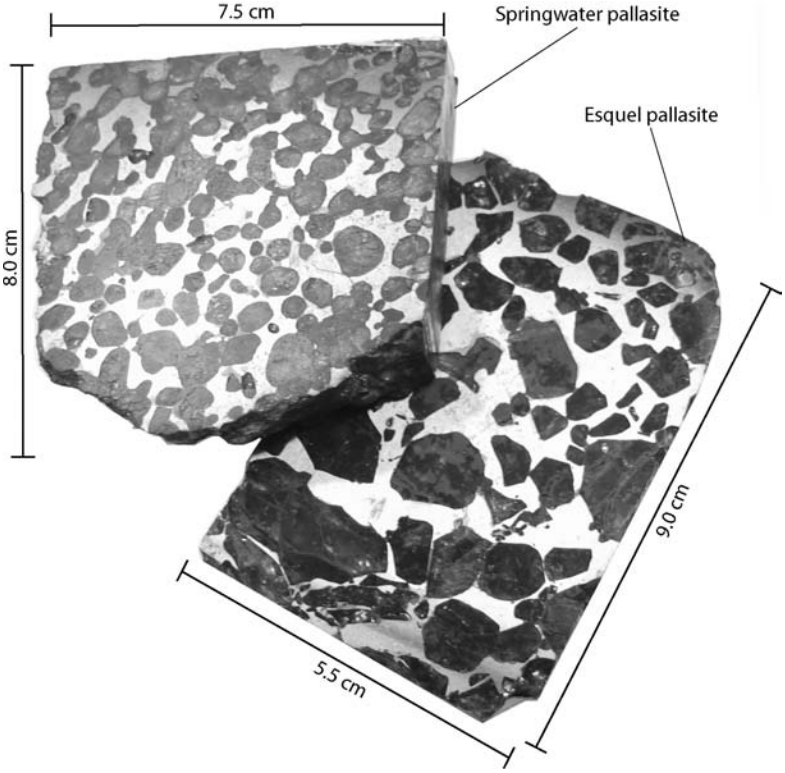


FIGURE 2.1 Cross sections of two pallasites (stony iron meteorites), by courtesy of J. Wasson and A. Rubin. The material with metallic appearance is troilite (FeS), which has melted and separated from the silicate.

and appearances quite different from terrestrial rocks. The characteristic feature is the presence of chondrules, finely crystalline but originally apparently glassy silicate globules up to a centimetre or so in size, that are embedded in a matrix of other materials. The word chondrite simply means a meteorite containing chondrules. For the ordinary chondrites, the matrix in which the chondrules are embedded is a mixture of crystalline silicate and, commonly but not necessarily, grains or filaments of nickel–iron. Also of special interest are refractory inclusions rich in Ca and Al (CAIs), that are distinct from chondrules and may be the earliest condensates in the solar nebula. Figure 2.2(a) shows a section across a meteorite with three types of material in a single specimen and Fig. 2.2(b) is a cross-section of the chondritic inclusion in it.

Since the chondrules constitute major fractions of most chondrites, it is evident that a large proportion of meteoritic matter is in the form of chondrules. However, they are uniquely meteoritic. None has been found in any terrestrial rock, but they represent a stage in the evolution of the materials that formed the terrestrial planets, and have survived in the chondritic bodies simply because they have never been heated strongly enough to rework them into fully crystalline rocks. As pointed out in Section 1.11, the chondrules were magnetized as independent grains and were subsequently incorporated in larger bodies. They were, therefore, a primary condensate in the solar nebula. The process of magnetization remains conjectural, but it appears that, as independent particles in the solar nebula, chondrules were subjected to transient heating and even melting. They are presumed to have acquired thermoremanent magnetizations in this process.

Chondrites are classified according to their chemistry and the degree of metamorphism, that is, the extent to which their mineralogy and structure were changed by heat and pressure in their parent bodies. The metamorphism caused redistribution of elements by diffusion between minerals. Iron (and nickel) in particular

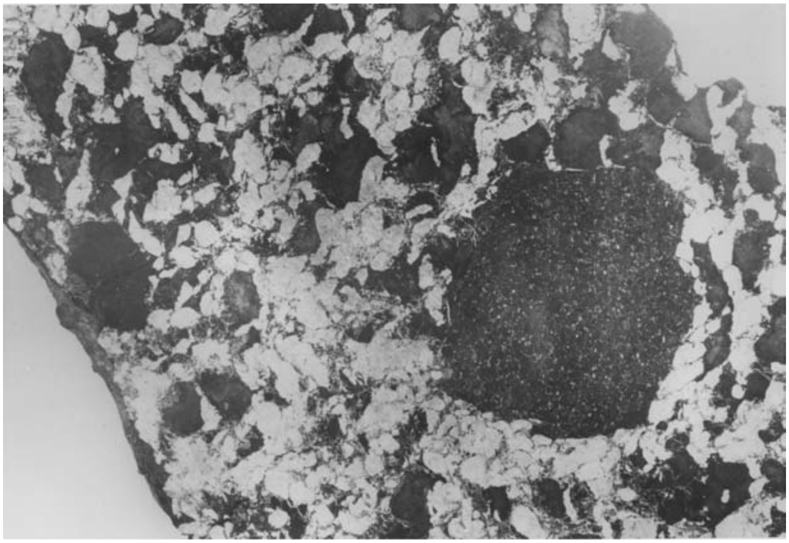


FIGURE 2.2(a) Polished section of the Bencubbin meteorite. The bulk of the meteorite has a well-developed crystalline (achondritic) structure, but a large chondrite inclusion appears in the dark area on the right-hand side of the photograph. On the left-hand side is a carbonaceous chondritic inclusion.

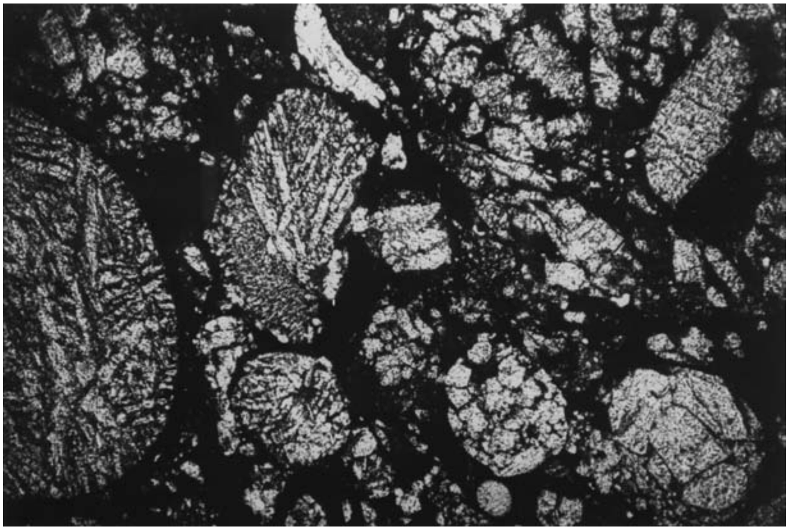


FIGURE 2.2(b) Enlarged section of the ordinary chondrite fragment, showing the structure of the chondritic spherules, just apparent in (a). Photographs courtesy of J. F. Lovering.

were redistributed in this way, as can be seen in the more strongly metamorphosed chondrites in which the metal grains tend to be either entirely taenite (high Ni) or entirely kamacite (low Ni) instead of being an intergrown mixture of both. But the taenite grains have the same diffusion profiles as the taenite lamellae in octahedrites (see Section 1.10 and Fig. 1.4), showing that they were effectively in contact with kamacite grains by diffusion of metal through the intervening silicate.

In classifying chondrites according to the degree of metamorphism, the most interesting are the least metamorphosed, because they are closest to the primitive material from which the terrestrial planets formed. These are the rare carbonaceous chondrites, so named because they contain carbon compounds. They are also much richer in volatiles than other meteorites and most of their iron is in the form of magnetite, Fe_3O_4. Carbonaceous chondrites are dark and amorphous in appearance and very friable. Their rarity in terrestrial collections is due more to their inability to survive flight through the atmosphere than to an absolute rarity in space. Many asteroids, especially those farthest out in the Solar System, have reflection spectra indicative of carbonaceous surfaces.

The carbonaceous chondritic material in museum and university collections is dominated by a single meteorite, Allende, which arrived as thousands of stones scattered over 300 km^2 in northern Mexico in 1969, following a spectacular fireball. Over 2000 kg were collected, and it is believed that this is only a small fraction of the total that fell. The fact that large sample sizes are available for analysis has permitted a range of experiments that would not have been possible without Allende. These include chemical analyses of large numbers of individual chondrules, and refractory inclusions (CAIs) of Ca-rich and Al-rich types, that suggest high temperature condensation from the solar nebula. These studies revealed anomalies in isotopic abundances, indicating that the nebula included dust grains from different nucleo-synthetic sources and that some of the grains survived in the most primitive meteorites (Section 4.5). But, except for light, volatile elements such as nitrogen and oxygen, the carbonaceous chondrites have compositions similar to that of the solar atmosphere and are believed to represent primitive solar nebular material. Isotopic studies of chondrites are discussed in Section 4.6.

2.5 Achondrites

Achondrites are stony meteorites that are fully crystalline, like terrestrial igneous rocks, and have no chondrules, although in cases such as the Bencubbin meteorite (Fig. 2.2) conglomerates occur. In the formal classification, achondrites have no metallic iron, but the stony parts of stony-irons (Fig. 2.1) are generally crystalline, like achondrites, so the classification is blurred. The achondrites are fragments of bodies that have evolved further towards planet formation than the ordinary chondrites. They are similar in composition to the mantle of the Earth. The evolutionary details presumably varied according to the sizes of the parent bodies and proximity to the Sun. A few of the achondrites have compositions representing further planetary evolution in bodies larger than asteroids. Meteorites collected in Antarctica include 30 or so that are identified as fragments of the Moon, thrown off by major meteorite impacts there, and one group has compositions indicative of Mars as the source. They include trapped gas, with nitrogen and the rare gases closely matching the proportions in the Martian atmosphere (see McSween, 1999, Fig. 5.19). They are known as SNC ('snick') meteorites from the initials of three representatives (Shergotty, Nakhla and Chassigny).

2.6 The solar atmosphere

Astrophysical estimates of solar abundances of elements are presented in terms of numbers of atoms, n, usually normalized to the number of hydrogen atoms, n_H. The standard form of presentation is

$$\log_{10} A = \log_{10}(n/n_\mathrm{H}) + 12, \qquad (2.1)$$

with the 12 chosen for convenience to make all the log A positive. When comparisons are made with meteorites or the Earth, normalization in terms of the number density of silicon atoms, n_Si, is preferred and the conversion presents no difficulty if we accept that

$$(\log_{10} A)_\mathrm{Si} = \log_{10}(n_\mathrm{Si}/n_\mathrm{H}) + 12 = 7.55 \qquad (2.2)$$

is a well-determined quantity for the solar atmosphere. In Table 2.1 silicon is used as the reference, but all values have been converted to proportions by mass, using the atomic masses, as listed in Table A.3 (Appendix A). Table 2.1 lists also the relative abundances in meteorites. This is an average for all meteorites and so may be biased against the carbonaceous chondrites, which are rare in terrestrial collections.

Recognizing the significance of carbonaceous chondrites, Ringwood (1966) calculated the composition that would result from heating these chondrites in a reducing (hydrogen) atmosphere to drive off volatiles, including CO, and produce metallic iron. His numbers agree rather well with the meteorite average, justifying the idea that aggregation of material in the inner Solar System began with accretion of carbonaceous chondritic bodies.

2.7 The mantle

The abundances of elements in the mantle in Table 2.2 are taken from McDonough and Sun (1995) with the addition of a value for oxygen, obtained by assuming that all of the elements except S occur as oxides. This may leave slightly too little for 'others', but any correction would require only a minor rescaling of the list. The table gives only those elements with abundances sufficient to require them to be taken into account in a density calculation. There are also some minor components that are very important to the behaviour and properties of the mantle. The thermally important elements, K, U and Th are considered in Chapter 21. Hydrogen, present as H_2O, OH^- ions and possibly very diffusive H^+ ions, is the most important of the volatiles that strongly influence rheology (Fig. 2.3). By comparison with the crust (Section 2.9) the mantle appears rather simple, as a consequence of the gleaning into the core of a range of siderophile elements and into the crust of elements described as 'incompatible', meaning that they do not fit into the major mantle minerals and concentrate in melts that rise to the surface. Although water is incompatible in this sense, it has an important role in the mineralogy and properties of the deep mantle (Ohtani *et al.*, 2001; Kombayashi *et al.*, 2005), either directly or as dissociated H^+ and OH^- ions.

Oxygen is by far the most abundant element in the mantle, even by mass, although it is the lightest element listed in Table 2.2. In terms of atomic numbers the dominance of oxygen is even more striking; more than 58% of the atoms are oxygen. Combined with the fact that O^{2-} is a large ion, mantle minerals can be viewed as lattices of O^{2-} ions with interstitial Si^{4+}, Mg^{2+}, Fe^{2+}, etc. in different proportions in the various minerals. The other fundamental control on mantle mineral structures is the strength of Si-O bonds and the strong preference of Si^{4+} ions for tetrahedral bonding. In the favoured crystal structures at ordinary pressures we see Si^{4+} ions at the centres of tetrahedra with O^{2-} at each of the corners. Mg^{2+}, Fe^{2+} and others occupy spaces between these tetrahedra. The resulting crystal structures are quite open with plenty of scope for closer packing at high pressure. Lower mantle pressures force the Si^{4+} ions into six-fold coordination with O^{2-}; in the most important of the lower mantle minerals, (Mg,Fe)SiO_3 perovskite, Si^{4+} ions occupy the centres of octahedra with O^{2-} at each of the six corners and Mg^{2+}, Fe^{2+} between the octahedra. The energy required for this change in Si coordination makes the phase transition at the 660 km boundary strongly endothermic, cooling the mantle minerals as they convert to the denser phase structure and opposing convection through that depth (Chapter 22).

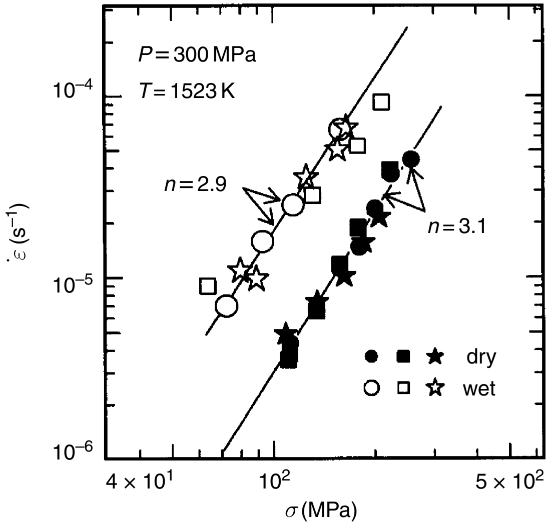


FIGURE 2.3 A log–log plot of deformation rate vs stress for olivine under hydrous and anhydrous conditions. *n* is the index of a power law relationship, as in Eq. (10.27). (After Mei and Kohlstedt, 2000b.)

The pyrolite model of the mantle, mentioned in the preamble, is dominated by two minerals, olivine (Mg_2SiO_4) and pyroxene ($MgSiO_3$), each occurring as solid solutions with Fe substituted for some of the Mg. The Mg end members of the two solid solution series are known as forsterite and enstatite respectively. From a comprehensive tabulation of crystal structures and densities by Smyth and McCormick (1995) we see that their mean atomic weights and densities are very similar (forsterite $\bar{m} = 20.099, \rho = 3227\,\mathrm{kg\,m^{-3}}$; enstatite $\bar{m} = 20.078, \rho = 3204\,\mathrm{kg\,m^{-3}}$). They are therefore gravitationally indistinguishable, with no tendency to separate, and the difference from the upper mantle density (about $3400\,\mathrm{kg\,m^{-3}}$ extrapolated to zero pressure and 290 K) must be explained by Fe substitution and the presence of other minerals such as garnet or, in the uppermost 100 km, $MgAl_2O_4$ spinel. The pyrolite model matches quite well a rock type known as peridotite, which therefore serves as a plausible mantle sample. Peridotite is a combination of olivine with two pyroxenes, orthopyroxene and clinopyroxene, with slightly different structures and densities but the same basic chemical formula. Their coexistence is an indication of the subtlety of a mixture in which none of the minerals is a pure compound and they have different responses to additions of other elements. Clinopyroxenes are more variable in composition than orthopyroxenes.

With increasing pressure a third mineral type, garnet, becomes important. Pyrope is a garnet with the formula $Mg_3Al_2Si_3O_{12}$ and a mean atomic weight of 20.156. This is only marginally higher than for forsterite and enstatite, but the density, $3565\,\mathrm{kg\,m^{-3}}$, is much higher. Fe and Ca can both substitute for Mg so garnet is the obvious mineral type to accommodate Ca as well as Al, but the significance of its density is that it is favoured by pressure. It is probably rare in the uppermost 100 km of the mantle, where $MgAl_2O_4$ spinel occurs, but extends into the transition zone and possibly even into the top of the lower mantle. Thus the upper mantle has a mineral structure that is somewhat depth dependent even before the first of the major phase transitions that convert the most abundant minerals to denser forms.

The phase transitions in olivine are clearest and explain the seismologically observed boundaries. The successive crystal structures of forsterite, with densities extrapolated to zero pressure and 290 K, are listed in Table 2.4a with transition pressures and corresponding mantle depths. The density increments at the 410 km and 660 km boundaries are dominant. The transitions in pyroxenes are not as sharp and may not contribute to the observed boundaries. But, as in Table 2.4b, they follow a similar trend. Recognized in these tables is the discovery of a 'post-perovskite' phase (Murakami *et al.*, 2004), for which details of transition pressure and temperature must be regarded as preliminary. This transition is presumed to contribute to the observed heterogeneity of the D″ layer at the base of the mantle. Note that these tables give zero pressure densities of all crystal forms, to allow comparison of intrinsic densities, but

Table 2.4a Phase transitions in olivine, Mg_2SiO_4. Zero pressure, 290 K densities of the different crystal structures with equilibrium transition pressures and corresponding depths in the Earth

Crystal structure	ρ_0 ($\mathrm{kg\,m^{-3}}$)	$\Delta\rho_0$ ($\mathrm{kg\,m^{-3}}$)	P (GPa)	z (km)
forsterite	3327			
		246	13.7	410
β spinel (Wadsleyite)	3473			
		75	17.9	520
γ spinel (Ringwoodite)	3548			
		395	23.3	660
$MgSiO_3$ perovskite	4107			
+ MgO periclase	3583			
	(3943 together)			
		∼60	∼120	∼2600
'post-perovskite'				
+ MgO periclase	∼4004 (together)			

Table 2.4b Phase transitions in orthopyroxene, $MgSiO_3$

Crystal structure	ρ_0 (kg m^{-3})	$\Delta\rho_0$ (kg m^{-3})
enstatite	3204	
		309
garnet	3513	
		297
ilmenite	3810	
		297
perovskite	4107	
		∼100
'post-perovskite'	∼4200	

that the high pressure forms are only metastable at zero pressure.

The lower mantle is dominated by perovskite and magnesiowustite (ferropericlase), with a greater concentration of Fe in the magnesiowustite and perovskite taking up the Al. Neither accepts Ca, and a few percent of $CaSiO_3$ perovskite is believed to occur as a third mineral. In total mass the (Mg-Fe) perovskite must be dominant, comprising 75% to 80% of the lower mantle, making it the most abundant mineral in the Earth. This has prompted both experimental and theoretical studies of its properties. It can be produced in a metastable state at zero pressure, but it does not withstand more than very limited heating so that elastic moduli are well observed (Yeganeh-Haeri, 1994) but thermal properties less so. Periclase (MgO) is stable over the whole range of temperatures and pressures of interest and its properties are well documented. The Ca perovskite does not survive decompression and can be studied only in high pressure experiments.

Justification is needed for a mantle composition that is in imperfect agreement with the elemental abundances of carbonaceous chondrites. They are not so rare in terrestrial collections that we can appeal to inadequate sampling. The differences in elemental abundances require there to have been systematic variations with radius in the solar nebula. Si is a particular problem, with relatively more Si at asteroidal distances than at the Earth's distance. Allègre *et al.* (1995) prefer to suppose that the missing Si is in the core but we do not favour this for reasons considered in the following section. We know that the density of Mercury requires a much higher proportion of iron than either the Earth or the meteorites, so there can be no compelling reason for rejecting heterogeneity of the nebula, and it is not difficult to find reasons for it, such as selective centrifuging of ionized atoms by the early solar magnetic field. However, this means that some independent observations are required to give confidence that the mantle composition has been correctly assessed. Rock samples that are inferred to come from the mantle, such as fragments brought up with volcanic magma (xenoliths) and peridotite nodules in kimberlites (diamond-bearing plugs of deep volcanic origin), offer strong circumstantial evidence, but it is difficult to be certain that they have not been modified on the way up. In any case evidence of the upper mantle composition does not answer a crucial question: how near is this to the lower mantle composition? The most convincing evidence comes from minute inclusions in diamonds of evident lower mantle origins (Kesson and Fitzgerald, 1992). These are grains of enstatite, that would have formed by decompression of lower mantle perovskite, mixed with magnesiowustite, just what is expected for lower mantle mineralogy, and consistent with a bulk composition similar to that of the upper mantle.

2.8 The core

There are almost certainly many elements dissolved in the core. Siderophile (iron-loving) elements that must be concentrated there include Ni, Co, Re, Os, Pt and Pd, but all of these are more dense than iron and, with the exception of Ni, are not sufficiently abundant to include in a density calculation anyway. Poirier (1994) reviewed the rival suggestions for light additives to iron that would reduce its density to that of the core. The first step in calculating how much of them is required is to determine the density deficit. We use an equation of state study (Stacey and Davis, 2004) that gives densities of pure iron in the ϵ (hexagonal close-packed) form that is stable at high pressures, extrapolated to zero pressure and 290 K ($8352 \pm 23\,\mathrm{kg\,m^{-3}}$), and outer core material solidified to the same structure, cooled and decompressed to the same state ($7488 \pm 30\,\mathrm{kg\,m^{-3}}$). The difference is 10.3% of the pure iron density, in close agreement with an early estimate by Birch (1952). Before accounting for this in terms of light elements, we allow for an increase in density due to Ni. This is chemically very similar to iron and forms simple substitutional alloys, so that, for the modest concentration considered, we can take its density effect to be directly proportional to atomic weight.

The average Ni/Fe ratio in chondrites of all types, as listed by McDonough and Sun (1995), is about 0.057. This is probably the appropriate ratio for the mantle but is too small to have produced the Widmanstätten exsolution patterns in iron meteorites, as seen in the examples in Figs. 1.3 and 1.5. We consider that the iron meteorites are likely to be a better approximation to the core composition than the chondrites and use a histogram of the Ni contents of iron meteorites by McSween (1999, Fig. 6.2) to estimate $\mathrm{Ni/(Fe+Ni)} = 0.082$ (by mass) for the core. On this basis, the core alloy, without light ingredients, would have a mean atomic weight $\bar{m} = 56.07$. In the high pressure (epsilon) form, the density, extrapolated to zero pressure and 290 K, would be 8385 $\mathrm{kg\,m^{-3}}$, making the core density deficit to be explained by light elements 10.7%.

As mentioned in Section 2.1, the favoured light elements are H, C, O, Si and S. Selection from these of a mixture that best explains the core density depends on what is assumed about accretion of the Earth and formation of the core. Thus H and C are abundant in carbonaceous chondrites and would be strong candidates if the Earth accreted from carbonaceous material, with subsequent chemical reaction to produce iron in a high pressure environment rich in these elements. We prefer to suppose that the nebular material was pre-processed and that most of the planetesimals from which the Earth accreted resembled iron meteorites and achondrites, which had formed at low pressures. Then, if core separation occurred with iron and silicate more or less in chemical equilibrium, we can use the rather low H and C abundances in the mantle to argue that, even with strong partitioning into iron, the core content of these elements must be modest. High pressure experiments by Okuchi (1997, 1998) make a strong case for partitioning into the core of such H as was available but the probable lower mantle content of H_2O suggests only about 4 atomic% (0.08% by mass) in the core. This number is assumed in Table 2.5. Similarly Wood (1993) argued that at least some C must have found its way into the core and this is also allowed for in Table 2.5 but, by our estimate these two elements together account for only about 10% of the density deficit. The core is mainly liquid, but with a solid inner core that has 5% of the total mass. Its seismologically estimated density contrast is 820 $\mathrm{kg\,m^{-3}}$ (Masters and Gubbins 2003), of which only 200 $\mathrm{kg\,m^{-3}}$ is explained by solidification. On this basis the density deficit of the inner core is 5.9%.

Table 2.5 Effect on core density of elements added to iron

Element	Ni	H	C	O	S	Total
$(\rho_{Fe}/\rho^* - 1)$	−0.049	7.93	1.15	0.95	0.66	
Vol./atom[a]	1.00	0.16	0.46	0.56	0.95	
Outer core						
mass %, f	6.49	0.08	0.50	5.34	8.44	20.85
$f\,(\rho_{Fe}/\rho^* - 1)$	−0.0032	0.0063	0.0057	0.0507	0.0557	0.1153
Inner core						
mass %, f	6.92	0.07	0.45	0.11	8.02	15.57
$f\,(\rho_{Fe}/\rho^* - 1)$	−0.0034	0.0056	0.0052	0.0010	0.0529	0.0613

[a] Relative to iron atoms

Braginsky and Roberts (1995, Appendices D and E) compared the cases for O, Si and S as candidate light elements in the core. They pointed out that Si and S would be almost equally soluble in solid and liquid Fe under core conditions, but that O would strongly partition into the liquid. This is supported by calculations of Alfè *et al.* (2002). Thus, to explain the density contrast between the inner and outer cores it is necessary to assume a substantial outer core oxygen content, with very little in the inner core. Then, with abundant O in the outer core, Si is disallowed as a major constituent. If accretion and core separation had occurred under sufficiently reducing conditions to introduce anything like 10% of elemental Si then there would have been no O. So, the remaining core constituent must be S, with only mild partitioning between solid and liquid. Gessman and Wood (2002) reported that O dissolves more readily in Fe if S is also present. With these arguments the abundances in Table 2.5 are quite well constrained, although the relegation of H and C to minor roles invokes the assumption that the Earth accreted from processed meteoritic material.

The Ni, H and C abundances in the core are discussed above, and we assume slight partitioning of H and C between solid and liquid, with no partitioning of Ni, that is Ni/(Fe + Ni) = 0.082 in both outer and inner cores. We now estimate the abundances of O and S. Since we are working with proportions of elements by mass, densities add as reciprocals, so that with mass fractions $f_1, f_2, \ldots$ of additives to Fe the density is

$$1/\rho = f_1/\rho^*{}_1 + f_2/\rho^*{}_2 + \cdots + (1 - f_1 - f_2 - \cdots)/\rho_{\mathrm{Fe}}, \tag{2.3}$$

where ρ^* is the effective density of a constituent in dilute solution in Fe and ρ_{Fe} is the undiluted density of Fe. Values of ρ^* can be calculated from densities of Fe-H by Okuchi (1997, 1998), Fe-C by Ogino *et al.* (1984), with Fe-O and Fe-S densities discussed by Braginsky and Roberts (1995) and Alfè *et al.* (2002). For Ni, $\rho^*/\rho_{\mathrm{Fe}} = 1.051$ is taken to be the ratio of atomic weights. It is convenient to multiply Eq. (2.3) by ρ_{Fe} and deal with density ratios that are assumed to be independent of pressure, where only low pressure data are available. Then the equation can be rewritten

$$\begin{aligned}(\rho_{\mathrm{Fe}}/\rho - 1) = {}& f_{\mathrm{Ni}}(\rho_{\mathrm{Fe}}/\rho_{\mathrm{Ni}}{}^* - 1) + f_{\mathrm{H}}(\rho_{\mathrm{Fe}}/\rho_{\mathrm{H}}{}^* - 1) \\ & + f_{\mathrm{C}}(\rho_{\mathrm{Fe}}/\rho_{\mathrm{C}}{}^* - 1) + f_{\mathrm{O}}(\rho_{\mathrm{Fe}}/\rho_{\mathrm{O}}{}^* - 1) \\ & + f_{\mathrm{S}}(\rho_{\mathrm{Fe}}/\rho_{\mathrm{S}}{}^* - 1)\end{aligned} \tag{2.4}$$

with values of $(\rho_{\mathrm{Fe}}/\rho^* - 1)$ for each element listed in Table 2.5. The effective volumes per atom are also listed. In calculating abundances, the partition ratios for concentrations of O and S in solid vs liquid are taken as 0.02 for O and 0.95 for S, so that most of the density contrast between inner and outer cores is attributed to O. The percentages by mass of these elements are given in Table 2.5 for both inner and outer cores.

The association of S with Fe in meteorites, commonly occurring as troilite (FeS), as in the example in Fig. 2.1, makes the inclusion of S in the core appear inevitable. At low pressure S dramatically lowers the melting point of Fe, facilitating the separation of a liquid core. A long standing suggestion that potassium (K) is associated with S in the core has received close attention because radiogenic heat from ^{40}K offers a solution to the problem of core energy (Sections 21.4 and 22.7). Experiments on the partitioning of K between Fe-S and silicate liquids at high pressure (Gessman and Wood, 2002; Murthy *et al.*, 2003; Hirao *et al.*, 2006; Hillgren *et al.*, 2005; Bouhifd *et al.*, 2007) lead us to conclude that some K probably entered the core, although much less than some reports have suggested. A reason for differing estimates is that the partitioning is strongly affected by the presence of other elements as well as temperature. Gessman and Wood (2002) reported that the presence of alumina in their pressure vessel inhibited the uptake of K by Fe-S, but Bouhifd *et al.* (2007) used sanidine, $KAlSiO_3$, as the silicate in their experiments with no apparent inhibition by the presence of Al.

Most partitioning experiments have sought the equilibrium between metal and silicate, with both molten, but core separation would have begun as soon as the first molten iron appeared, before complete accretion of the Earth or formation of a magma ocean. The process would have begun at a temperature as low as 1500 K, but may have gone to completion only when the deep mantle reached 4000 K. With the temperature variation of the coefficient for partitioning of K between Fe-S and silicate reported by Bouhifd *et al.* (2007), its equilibrium concentration in the metal would have varied from 0.01 to 0.5 of the concentration in silicate. The core concentration is presumed to be somewhere in this range. The ratio in our thermal model (Table 21.3) is 0.4. Uranium may not have been securely discounted as a core constituent, but we follow the majority view, expressed by Wheeler *et al.* (2006), that it is unlikely to be significant.

The physical case for core radioactivity arises from the energy requirement of the geomagnetic dynamo, which is most easily satisfied if the solidification of the inner core is slowed by an additional heat source. However, the argument depends critically on the core energy loss by thermal conduction, and therefore on the conductivity, which is not well determined (Stacey and Loper, 2007). Our conductivity estimate (see Sections 19.6, 22.4 and 22.7) indicates a modest abundance of K in the core, sufficient to give 0.2 terawatt of radiogenic heat at the present time. This requires 29ppm of K in the core, 40% of its concentration in the mantle, as listed in Table 21.3. A much greater concentration appears implausible and it is still possible that the core has no radioactivity, although that requires a thermal conductivity lower than our estimate.

The core is believed to be cooling only slowly, but any cooling means progressive growth of the inner core by freezing of outer core liquid with rejection of the oxygen, which remains in the liquid as a source of buoyancy at the inner core boundary. This is an energy source for outer core convection. However, provided it is only O that partitions strongly into the liquid, with virtually none entering the solid, the increasing outer core concentration causes no compositional gradient in the inner core.

2.9 The crust

A plot of the distribution of elevations of the solid surface of the Earth is known as the hypsographic curve (Fig. 9.4). There are much larger areas close to sea level and at depths of 4 to 5 kilometres than at intermediate levels. Most of the crust is either of continental type or ocean basin type and the two are structurally very different. The mantle underlies both continental and oceanic areas, and is identified by a seismic P-wave speed of about 8.0 km/s, which is essentially the same everywhere. The crustal thickness of the ocean basins is about 7 km, including sediments but not the depth of sea water, whereas the thickness of the continental crust averages 39 km, with a maximum of 65 km+ under the Himalaya. The crust–mantle boundary, the Mohorovičić discontinuity, colloquially abbreviated to Moho, is sufficiently clearly observed by seismology to establish that it is distinct everywhere, except at mid-ocean ridges. The crust is a veneer differing in composition from the much greater mass of the mantle beneath it. The crustal structures in continental and oceanic areas are very different and the difference is central to our understanding of tectonics (Chapter 12). But neither the continental nor ocean basin crusts are uniform with depth and we start with a simplistic view by considering the upper layer of each.

The continental crust is an evolutionary product of the Earth over most or all of geological time. Its development by differentiation from the mantle would initially have been rapid, but continues to the present time. It is continuously recycled and modified by erosion–sedimentation and metamorphic and organic processes, and this is reflected in its complexity and diversity. The crust of the ocean floor appears much simpler. It is comparatively short-lived, being produced volcanically at ocean floor ridges and disappearing back into the mantle at subduction zones after a period of order 100 million years, only a few per cent of the ages of the oldest continental rocks. Ocean floor sediment and entrained sea water are carried down with the subducted crust-upper mantle layer and provide a flux for the development of Si-rich lavas that become continental crust. This is an essential feature of the recycling process that maintains the continental crust.

Representative igneous rocks found in the crust, in order of increasing SiO_2 content and decreasing ($MgO + FeO$), are listed in Table 2.6. The last three have compositions characteristic

Table 2.6 Average compositions of representative igneous rocks (per cent by mass)

	SiO_2	MgO	FeO + Fe_2O_3	Al_2O_3	CaO	Na_2O	K_2O
Komatiite	45.5	20.6	13.2	9.2	8.6	0.8	0.02
Eclogite	46.2	13.7	11.1	15.8	9.8	1.6	0.4
MORB[a]	47.5	14.2	9.5	13.5	11.3	1.8	0.06
OIB[b]	49.4	8.4	12.4	13.9	10.3	2.1	0.4
Andesite	59.2	3.0	6.9	17.1	7.1	3.5	1.8
Granite	72.9	0.5	2.5	14.5	1.4	3.1	3.9
Rhyolite	74.2	0.3	1.9	14.5	0.1	3.0	3.7

[a] Mid-ocean ridge basalt
[b] Ocean island basalt

of the continental crust, being acid, meaning SiO_2 rich, rocks. Andesite is a direct product of subduction zone volcanism. Rhyolite is also volcanic, but clearly more acid and is presumed to be recycled continental crust. The origin of granite is a subject of debate. It occurs as massive, and apparently very slow, intrusions with assimilation of the intruded rocks. The compositional similarity to rhyolite suggests a similar ultimate source and they may differ only in the degree of reheating and speed of cooling. The processes of recycling of continental rocks that produce them are not well understood.

As is obvious from the composition of granite, the dominance of SiO_2 ensures that, when all of the other oxides are combined with it as silicates, there is still plenty ‘left over’ to crystallize as quartz, which may be nearly pure SiO_2. The other minerals are mainly feldspars, with compositions such as $(Na,K)AlSi_3O_8$, and plagioclase, $CaAl_2Si_2O_8$, in which various substitutions occur. Granite is fairly coarsely crystalline, making the different mineral grains obvious in a freshly exposed section or hand sample. A consequence is that erosion and sedimentation can allow sorting of grains by density or grain size, under the actions of river flow, shoreline waves or wind, and leading to local concentrations of particular minerals. This may be a first stage in the development of exploitable deposits (almost the final stage in the case of beach sand titania and zircon). Metamorphic processing by heat, pressure and hydrothermal circulation, by which mineral constituents are dissolved in hot, percolating water and deposited elsewhere, modify crustal rocks, producing an amazing range of minerals (see, for example, Smyth and McCormick, 1995).

Two types of basalt, identified as MORB and OIB in Table 2.6, are distinguished by the depths of their mantle sources. MORB is alkali basalt, produced at mid-ocean ridges by partial melting of the upper part of the mantle, within about 100 km of the surface, and is regarded as an indicator of the upper mantle composition. This is more depleted in the incompatible elements that do not fit well into mantle mineral structures than the deeper sourced OIB, apparently because these elements have been gleaned from the upper mantle by earlier convection more completely than from the lower mantle. The OIB type of basalt is identified as partial melt from deep mantle plumes that carry core heat up through the mantle, although it is possible that there are also shallower sources of similar material and some modification on the way up is probable. It is composed of pyroxenes, minerals based on the $MgSiO_3$ structure, and plagioclase, $CaAl_2Si_2O_8$, commonly with olivine, $(Mg,Fe)_2SiO_4$ and some glass, indicative of rapid cooling. Alkali basalts, often with more Na and K than the MORB composition in Table 2.6, also include alkali feldspars, $(NaK)AlSiO_3$, or feldspathoids with the same elements in different proportions and crystal structures. Soils derived from weathering of basalt are generally very fertile. They are characteristically red soils, coloured by iron which is oxidized to hematite, Fe_2O_3. By comparison, soils derived from decomposed granite are less fertile.

There are systematic variations with depth of both oceanic and continental crusts. Marine geophysicists refer to layers 1, 2, 3 for the ocean floors, with seismic reflections from the boundaries between them commonly observed. Layer 1 is simply sediment, referred to below. Layer 2 is typically 1.5 km thick, with a P-wave speed of about 5.1 km/s, and is interpreted as familiar extruded basalt (MORB) affected by circulation of sea water through pores and cracks. Layer 3 is about 5 km thick with a P-wave speed of 6.7 km/s, with fewer pores and cracks and presumably weaker (or non-existent) hydrothermal circulation. It is more coarsely crystalline because of slower cooling. But the P-wave contrast with layer 2 demands also some compromise with the mantle and not just a deeper layer of uncracked MORB.

When we look at the continents we see granite as a typical component. But the abundance of heat-producing elements in it disallows consideration of a granitic layer extending to the Moho because that would provide more than the observed surface heat flux. Seismic reflections indicate complex structures. The deeper rocks must be more basic than the widespread granitic layer. The crustal layering of both continents and ocean floors indicates a separation of components in a melt or partial melt as igneous crust is forming, and that the shallow crust differs from the mantle more than do the deeper layers.

Erosion of the continents produces a flux of sediment to the oceans by river flow, estimated to be about 22×10^{12} kg/year (McLennan, 1995), but probably no more than half this in pre-agricultural times. About 80% is deposited on submarine margins of the continents and in estuaries and coastal wetlands, with about 4×10^{12} kg/year carried to the deep ocean basins. Its slow deposition is accompanied by precipitates of biological origin, especially $CaCO_3$, but since the calcium in sea water is dissolved from eroded rocks we can consider all of the deep sea sediment to be of continental origin. Its total mass is about 2.8×10^{20} kg. Dividing these numbers we see that the observed sediment would accumulate in 70 million years. This is a measure of the average duration of the ocean floor between its origin at a spreading ridge and return to the mantle at a subduction zone. Much, perhaps most, of this sediment is carried down with the subducting lithospheric slabs, as demonstrated by the ^{10}Be contents of andesitic lavas (Morris *et al.*, 1990). These authors also point out that boron is more abundant in andesites than can be explained by a mantle source and that it originates in sea water carried down with the sediments. The particular significance of ^{10}Be is that is that it is a radioactive isotope with a half life of 1.5×10^6 years, produced by cosmic ray bombardment of the upper atmosphere, washed into the sea and deposited with the sediment. Its existence in andesitic lavas means that the interval between subduction and volcanic re-emergence is not many multiples of the half-life, certainly less than 10 million years.

The mass balance of continental erosion, sedimentation and recycling compels the conclusion that most, if not all, of the sediment carried into the sea is returned to the continents by reworking and underplating, as well as volcanism (Section 5.3). The diversity of continental igneous material indicates a complex history in which sedimentation has a central role. It selects and redistributes minerals, so that when they are reheated and compressed they re-emerge as a variety of igneous rock types.

2.10 The oceans

Sea water contains 3.5% by mass of solutes, listed in Table 2.7. The solute concentration is locally variable by 10% of this value, but the proportions of the major elements are very consistent. The mixing of sea water by its circulation is very rapid compared with fresh input or removal of solutes and only the minor constituents linked to biological cycles or human activity vary with depth or season. Sea water is slightly alkaline, represented by a pH of 8, controlled primarily by a continuous exchange of CO_2 with the

Table 2.7 Solutes in sea water as parts per million by mass of elements. From Fegley (1995)

Element	Abundance (ppm)
Cl	19 353
Na	10 781
S as sulphate	2712
Mg	1280
Ca	415
K	399
Br	67
C as CO_2	26.4
N as N_2 gas	16.5
as nitrate	0.84
Sr	7.8
O as O_2 gas	4.8
B	4.4
Si as silicate	3.09
F	1.3
U	0.0032

atmosphere in a balance with carbonate, $(CO_3)^{2-}$, bicarbonate, $(HCO_3)^-$ and Ca^{2+} ions. The total CO_2 dissolved in the oceans is about 20 times that in the atmosphere.

It was at one time supposed that the rate of transport by rivers to the sea of NaCl dissolved from eroding rock gave a measure of the age of the Earth. However, the exchange with the solid Earth is much more complicated and not well constrained by observations. There is an exchange of solutes with the crust by hydrothermal circulation of sea water through cracks near the axes of spreading centres (mid-ocean ridges). This is apparent from the accumulation of hot brine in hollows on the floor of the Red Sea, a nascent mid-ocean ridge (Degens and Ross, 1969), where there is no effective deep ocean circulation.

The oceans are the major reservoir of the Earth's water, but not the only one, and the several reservoirs are all linked. Exchange with the atmosphere is most obvious. About 25% of the rain water falling on land flows to the sea in rivers, but much of it is directly re-evaporated or transpired by vegetation, and the balance sinks in to maintain the store of ground water that leaks more gradually to the sea. Most natural lakes are windows to the water table and the effectiveness of bores for the supply of water indicates a massive global store of it. Of more particular interest to the theme of this text is the deep exchange of water with the solid Earth and its role in controlling properties of rocks and minerals. This is a subject of the following section.

2.11 Water in the Earth

Water is only a minor constituent of the Earth as a whole, although it is abundant at the surface. Its physical and chemical properties give it a controlling influence on our environment. It occurs in all three phases (solid, liquid, gas) and the latent heats of melting and evaporation are essential to the redistribution of heat over the surface. Water is one of very few materials that expand on freezing, allowing liquid water to remain underneath a frozen surface. In fact, the expansion by cooling begins above the freezing point; the thermal expansion coefficient is negative for very cold water. The temperature of maximum density is 4 °C for pure water and 2 °C for sea water, so that cold polar water, still safely above freezing point, sinks to the sea-floor and flows over all the ocean floors, maintaining a uniform, constant temperature. It completes a cycle of ocean circulation that carries equatorial heat polewards. The isothermal ocean floor makes possible the estimation of sea-floor heat flux from the temperature gradient in the upper few metres of sediment (Chapter 20). Another crucial fact is that the water molecule is lighter than the other atmospheric gases, so that its evaporation from the surface stimulates atmospheric convection and consequent cycling of water through the atmosphere.

An isolated oxygen atom has filled 1s and 2s electron states and four electrons in the six available 2p states, which, unlike the s states, are asymmetrical. In water the p states are shared with the electrons of hydrogen in bonding that is partly covalent but partly ionic, so that the oxygen and hydrogen atoms are oppositely charged. The asymmetries of the interacting p states make the molecular structure asymmetrical, with O-H bonds oriented at an angle, so that the negatively charged oxygen is displaced from the point mid-way between the H^+ ions, giving the H_2O molecule an electric dipole moment. This is responsible for many of the properties of water. It is a good solvent for polar molecules, such as NaCl, which dissociates into Na^+ and Cl^- ions that are attracted to the opposite charges of the water molecules, making the solution an electrical conductor. Very few of the H_2O molecules dissociate in pure water; it is solutes that give water the reputation of being a conductor. But ground water always has enough solutes to make it conducting for the purpose of electromagnetic exploration and sea water is sufficiently conducting to screen the sea-floor from rapid geomagnetic disturbances.

We now re-examine the role of water within the Earth. As mentioned in Section 2.9 and Chapter 12, water in interstices and hydrated minerals in ocean floor sediments is carried down with lithospheric material in subduction zones, locally lowering the solidus temperature (at which partial melting occurs) and leading to andesitic volcanism. There are no direct observations of the balance between subducted water and the water released to the atmosphere in volcanos, but most of it is presumed to partition into the magma and not to have a permanent effect on the water content of the mantle. There is generally less water in MORB than in OIB, consistent with its classification with the 'incompatible' elements that are gleaned into the crust by volcanism, and are more depleted in the upper mantle than in the less processed lower mantle. The solid Earth is probably continuing to lose water slowly. However, the rate is far from sufficient to accumulate the oceans in the life of the Earth and they must have been established early. But the water contents of basaltic lavas that have not acquired subduction zone water (as have andesites) suggest that the water still remaining in the mantle is comparable to the water of the oceans. This means that it is not changing very significantly and that mechanical properties that are influenced by it are sensibly constant; it does not need to be treated as a variable in calculations of thermal history (Chapter 23).

Free water is known to lubricate faults and to release earthquakes that would not occur under dry conditions but it can exist only to moderate depth, possibly limited to the upper crust. Hydrated minerals which structurally incorporate water are well known, but they too, probably have a limited depth range and more important at depth would be minerals with structures including $(OH)^-$ ions (see for example the list of mineral structures by Smyth and McCormick (1995)). But such minerals would not account for the phenomenon of hydrolytic weakening, which is important to mechanical properties and requires widely distributed $(OH)^-$ and/or H^+ ions that would locate at crystal imperfections in the host minerals and should be regarded as interstitial. Since oxygen is ubiquitous this is equivalent to incorporation of water. The strength of rock is largely attributable to the strength and angular rigidity of Si-O bonds and the effect of interstitial $(OH)^-$ or H^+ is to provide alternative bonding, facilitating the breaking of Si-O bonds. Measurements by Mei and Kohlstedt (2000a, 2000b) of the rate of deformation of olivine at high temperature and pressure under hydrous and anhydrous conditions (Fig. 2.3) illustrate the weakening effect of water. The rheology of the Earth, as inferred from post-glacial rebound (Chapter 9) and mantle convection (Section 13.2), requires some water at all depths.

It remains to ask why water is not more evident on other planets. As documented by McSween (1999), carbonaceous chondrites contain up to 18% water, of which only a tiny fraction would be required for planetary oceans. Mars may once have had surface water that could have produced the features suggestive of erosion if kept liquid long enough. The ready escape of hydrogen from dissociated water in the Martian atmosphere would allow dissipation of the water if it could get high enough in such a cold atmosphere for ultra-violet exposure, but that leaves the question: what happened to the oxygen? It may have been consumed in oxidation of the crust. In the case of Venus, the very limited water in the atmosphere is not easily explained in view of its ability to retain light gases. Perhaps the startlingly high $^2H/^1H$ ratio holds a clue if that could be understood.

However, satellites of Jupiter have water and even indications of saline liquid oceans underneath deep-frozen capping (Kivelson *et al.*, 2000).

2.12 The atmosphere: a comparison with the other terrestrial planets

Selected data on the atmospheres of the three terrestrial planets large enough to hold them are presented in Tables 2.8 and 2.9. They are very different and give some surprises that need to be examined for clues to the evolutionary histories of the planets. The most obvious feature of Table 2.8 is the similarity in relative abundances of the major elements in the atmospheres of Venus and Mars and the great dissimilarity to the Earth. Venus and Mars are in many ways (size, proximity to the Sun and surface temperature) very different, so the atmospheric similarity suggests a composition close to the primordial one with which all the terrestrial planets started, dominated by CO_2 and N_2. Then we see the atmosphere of the Earth as having developed from this by biological activity and note the requirement for water. The basic underlying reason why the Earth is different is that it has surface water. As mentioned in Section 2.9 and Chapter 12, water is also responsible for the style of the tectonic processes of the Earth and the resulting outgassing further modifies the atmosphere. We note also the possible importance of another difference: the Earth has a magnetic field that protects the atmosphere from direct exchange with the solar wind.

In comparing the numbers in Table 2.8 it must be noted that these are relative abundances and that the atmospheric densities of Venus and Mars differ by a factor exceeding 400. Relative to the planetary masses, all constituents except oxygen and argon are more abundant on Venus. Total abundances by mass, relative to planetary masses, are listed in Table 2.9 for three isotopes that appear particularly significant. This table

Table 2.8 Atmospheres of terrestrial planets: abundances of constituents (parts per million by volume) and some relevant properties. A variable water content is added to the Earth's atmosphere

Constituent	Venus	Earth	Mars
N_2	35 000	780 840	27 000
O_2	–	209 440	1300
Ar	70	9340	16 000
CO_2	965 000	364 (year 2000)	953 200
Ne	7	18	2.5
He	12	5.2	–
CH_4	–	1.7	–
Kr	0.025	1.14	0.3
N_2O	–	0.32	–
Xe	0.019	0.086	0.08
SO_2	185	5×10^{-5}	–
Properties			
Atm. mass/planet mass	1.01×10^{-4}	8.79×10^{-7}	3.9×10^{-7}
Mean mol. wt	43.45	28.97	43.34
Surface gravity ($m\,s^{-2}$)	8.87	9.78	3.69
Grav. potential ($10^7\ m^2\ s^{-2}$)	−5.369	−6.258	−1.264
Surface pressure (10^5 Pa)	95	1.01	0.064
Surface temperature (K)	737	288	215
Planet mass/Earth mass	0.815	1	0.107

Table 2.9 Atmospheres of terrestrial planets: some indicative ratios. Except for the last three entries the numbers refer to numbers of atoms or molecules and must be multiplied by atomic weights to obtain ratios by mass

Ratio	Venus	Earth	Mars
^{2}H/^{1}H	0.016	1.56×10^{-4}	8×10^{-4}
^{16}O/^{18}O	500	498.7	500
^{40}Ar/^{36}Ar[a]	1.1	296	3000
^{40}Ar/^{4}He[b]	5.8	1796	$\sim\infty$
Ne/Kr	280	16	8
Kr/Xe	1.3	13	4
CO_2/N_2	27.6	4.66×10^{-4}	35.3
Mass of ^{40}Ar/planet mass	3.4×10^{-9}	11.36×10^{-9}	5.8×10^{-9}
Mass of ^{36}Ar/planet mass	2.8×10^{-9}	3.5×10^{-11}	1.7×10^{-12}
Mass of ^{4}He/planet mass	5.9×10^{-11}	6.3×10^{-13}	–

[a] Ratio in solar wind 0.14
[b] Total production ratio in 4.5×10^{9} years ~ 0.13

also lists isotopic ratios. The ability of a planet to retain atmospheric gases is controlled by temperature and gravity, which are listed for each planet in Table 2.8.

Helium and free hydrogen escape from the atmospheres of all of these planets. This can be seen by comparing the ^{40}Ar/^{4}He ratios with the production of these isotopes in the life of the Earth (see footnote to Table 2.9). Assuming, for simplicity, similar degrees of outgassing for these isotopes, with complete retention of argon and similar ratios of radioactive elements, we see that Mars has retained no measurable He, the Earth has retained 7 parts per million and Venus has a remarkable 2% of the total He production. This is not what would be expected from its high temperature and weaker gravity than for the Earth. The difference appears greater than can be explained by diffusion-controlled escape from the dense atmosphere of Venus, and the only other obvious difference is that the Earth has a magnetosphere.

The Earth's atmosphere is believed to have retained the ^{40}Ar that has leaked into it over geological time and the very slight ^{36}Ar content that accompanied it is primordial gas that was trapped in the Earth when it formed. To explain the 100-fold greater ^{36}Ar content of the Venus atmosphere, we appeal to exchange with the solar wind, in which ^{36}Ar is the dominant Ar isotope. Then we can use the same explanation for the high ^{4}He and ^{2}He abundances in the Venus atmosphere. The only obvious reason for the big differences in the isotopic compositions, relative to the Earth, is that the Earth's atmosphere is protected from direct interaction with the solar wind by the magnetosphere. The assumption that this is so is necessary to the argument that the ^{40}Ar content of the Earth's atmosphere is a measure of the ^{40}K content of the Earth. The lower ^{40}Ar contents of the Venus and Mars atmospheres appear to suggest that those planets are less outgassed than is the Earth, but in view of the evidence for exchange between their atmospheres and the solar wind, it is possible that they have lost ^{40}Ar, especially so in the case of Mars. It is not clear that we can explain the abundances of the rare gases, Ne, Kr, Xe, in the same way. Ozima and Podosek (1999) pointed out that the abundance of Xe in the Earth's atmosphere appears to be anomalously low, and this is seen in the comparison with Kr in Table 2.9.

The dominant gases in the Venus and Mars atmospheres, CO_2 and N_2, have very similar proportions, inviting the conclusion that they approximate the primordial atmospheres and that the early atmosphere of the Earth was similar. Most of the Earth's CO_2 has been sequestered as $CaCO_3$ in the shells of marine organisms and fossilized as limestone. Some of the nitrogen has also probably been sequestered by biological activity and buried in the Earth. A unique feature of the Earth is its oxygen atmosphere. This is released by photosynthesis, with burial of carbon in reduced form as coal, oil and a much larger mass of less concentrated fossil carbon. This is the most important mechanism for generation of an oxygen-rich atmosphere, but not the only one. Dissociation of water vapour in the upper atmosphere by solar ultra-violet radiation releases hydrogen that may escape to space, leaving the oxygen gravitationally bound to the Earth. It is possible that this is the explanation for the small oxygen content of the Martian atmosphere. But the rate of loss of hydrogen from the Earth's atmosphere, as estimated from a Doppler shifted reflection from ionospheric hydrogen, is only about $0.2\,\mathrm{kg\,s^{-1}}$ and at this rate would have released oxygen amounting to no more than 20% of the atmospheric abundance. This is much less than sufficient to support the oxygen loss by weathering of crustal rocks. The biosphere is a much greater net producer of oxygen, but only by virtue of the burial and fossilization of reduced carbon, most of it probably carried down into the mantle in ocean sediment at subduction zones. Consumption of oxygen by decay processes would balance production without continuous removal of carbon.

3

Radioactivity, isotopes and dating

3.1 Preamble

Radioactive decays of certain naturally occurring isotopes are widely used to date terrestrial and meteoritic materials and to trace their evolution. Long before the discovery of radioactivity in 1896, it was understood that geological events occurred in a recognisable sequence, but attempts to fit them to a time scale were very insecure and contentious (see Section 4.2). Sedimentation and the fossil record are still central to geological history but now the fossil-based geological periods are linked to isotopically dated events. The principles of dating by radioactive decay require precise measurement of isotopic abundances. Isotopic methods have become so sensitive that very small variations in isotopic ratios of light elements, arising independently of radioactivity, are also routinely measured (Section 3.9).

We distinguish three categories of radioactive isotope that are of interest (Tables H.1, H.2, H.3 of Appendix H). Table H.1 lists the isotopes that are not produced in the Earth or the atmosphere by any continuing process, and must be accounted for in the inventory of elements in the Earth's original accretion. In only one important case (^{235}U) is the half-life less than 10^9 years and then only marginally so (a very rare isotope, ^{146}Sm, has a half-life of 10^8 years). Many shorter-lived species would have been produced at the same time but have now disappeared. This is a clue that the last of the nuclear synthetic events that produced the material of the Solar System occurred several billion years ago. The use of isotopes and radioactive decays to date the formation of the elements and the Earth was pioneered by Ernest Rutherford, whose work is documented in an illuminating review by Fowler (1961). Rutherford noted that for elements with even atomic numbers, z, which include uranium ($z = 92$), the isotopes with even atomic masses are normally more abundant than those with odd atomic masses. He concluded that ^{235}U was never as abundant as ^{238}U and, using the fact that ^{235}U has a much shorter half-life, imposed an upper bound on the age of these isotopes. With modern values of the half-lives and the ratio of present abundances, his argument imposes an age limit of 5.9 billion years. This is much less than the age of the Universe, as inferred from the Hubble constant and the cosmic microwave background radiation (13.7 billion years). But the lack of elements with half-lives much less than that of ^{235}U indicates that the Earth is not dramatically younger than these elements. As we now recognize, the interval between the synthesis of heavy elements and the formation of the Solar System was much shorter than the subsequent life of the Earth (Section 4.4). This is the reason for interest in some of the shorter-lived isotopes in Table H.3, because although no measurable amounts remain, they left decay products that are identified as 'orphans' in meteorites (Section 4.4) and provide more direct estimates of the synthesis–accretion interval.

There are also short-lived naturally occurring radioactive elements (Table H.2), but they are either produced by cosmic ray bombardment of the upper atmosphere, and precipitated with rain, or continuously produced in the Earth or oceans as intermediate daughters in the decay chains of uranium and thorium. They are used as tracers of geological processes with shorter time scales than those studied by the isotopes in Table H.1. One of the most interesting of the cosmic ray produced isotopes is ^{10}Be, which accumulates in marine sediment, disappears into the mantle at subduction zones and re-emerges with andesitic lava (Section 2.9). It demonstrates that wet marine sediment is subducted, becoming a flux for andesitic magma (Section 2.5), and that the whole process takes only a few ^{10}Be half-lives, less than 10 million years. The most useful example of an intermediate daughter isotope is ^{230}Th, a direct product of ^{234}U in the decay chain of ^{238}U, ultimately decaying to ^{206}Pb. ^{230}Th, with a half-life of 75 000 years, is produced in the shells of marine creatures that incorporate some uranium, and provides a dating tool for carbonate sedimentation.

Small variations in the relative abundances of isotopes of light elements arise from ordinary physical and chemical processes (Section 3.9), without radioactivity. Mass differences between isotopes cause mass-fractionations, so that, for example, water evaporating from the oceans is slightly depleted in deuterium relative to sea water, because light molecules evaporate more readily than the heavier ones. Partitioning of isotopes also occurs between interacting minerals and reflects the conditions (temperature and pressure) under which they come to equilibrium. More dramatic isotopic variations are found in fine grains in carbonaceous chondrites (Section 2.4), but are attributed to the preservation of unmixed material from different nucleo-synthetic sources. They present a clue to the pre-history of the material of the Solar System (Section 4.5).

Another reason for interest in radioactivity is that it is a source of heat. It is the dominant continuing energy source in the Earth (Chapter 21) and its distribution is central to the discussion of thermal history (Chapter 23). In this context there are four important isotopes, ^{238}U, ^{235}U, ^{232}Th, and ^{40}K. They are concentrated in the crust but are distributed throughout the mantle. The existence of radioactivity in the core has been contentious but, if there is some radiogenic heat, it eases the problem of finding an adequate energy source for the geomagnetic dynamo (Chapter 24). The case for some K in the core is discussed in Section 2.8 and the implications for thermal history in Chapter 23.

3.2 Radioactive decay

The rate of radioactive decay of an isotope is represented by the decay constant, λ, which is the probability per unit time that a constituent particle in an atomic nucleus will escape through the potential barrier binding it to the nucleus. Thus the rate of decay of N nuclei is proportional to N:

$$\frac{\mathrm{d}N}{\mathrm{d}t} = -\lambda N. \tag{3.1}$$

Integrating from an initial number N_0 at time $t=0$ we obtain the decay equation

$$N = N_0 \mathrm{e}^{-\lambda t}. \tag{3.2}$$

The relationship between λ and the half life, $\tau_{1/2}$, of an isotope is obtained by substituting $N = N_0/2$ at $t = \tau_{1/2}$,

$$\tau_{1/2} = \frac{\ln 2}{\lambda} = \frac{0.69315}{\lambda}. \tag{3.3}$$

Nuclear binding energies are so large and atomic nuclei are so small that radioactive decay is almost unaffected by physical conditions in the Earth, such as temperature and pressure. Decay by escape of α-particles (^{4} He nuclei) and β^- or β^+ particles (electrons or positrons) occurs by the penetration of potential barriers that bind these particles to the nuclei. The probability of escape is solely a property of a nucleus. The probability of decay by fission, in which a nucleus breaks into two comparable fragments plus neutrons, is also a nuclear property represented by a decay constant. A different process is the capture of orbital electrons. This is known as K-capture because almost always it is an electron from the innermost (K) shell of electrons that is captured. In this case the rate depends on the local density of orbital electrons at the nucleus. This is increased slightly by pressure, but much less than the density of solid material, which is controlled by the more compressible outer electron shells. Examples of K-capture include the decay of ^{40}K to ^{40}Ar. ^{40}K decays by three competing processes, β^- decay to ^{40}Ca (89.5%), K-capture to ^{40}Ar (10.5%), with a very small contribution by β^+ emission. Thus the rate of decay of potassium to argon in the deep interior of the Earth is probably very slightly greater than in the crust. This effect has not been measured for ^{40}K but is certainly small, and for practical purposes λ and $\tau_{1/2}$ for all isotopes, including ^{40}K, can be regarded as constants.

3.3 A decay clock: ^{14}C dating

A decay clock is one that uses Eq. (3.2). The measured abundance, N, of a decaying isotope is compared with an assumed initial abundance, N_0, and t is calculated from the ratio. The need to know N_0 restricts the usable decay clocks to those that make use of continuously maintained reservoirs of the parent isotopes (Table H.2). The most important of these is ^{14}C, which is produced by the (n, p) reaction of cosmic ray-generated neutrons on atmospheric ^{14}N. ^{14}C is incorporated in vegetation by photosynthesis, so that materials of biological origin can be dated by the ^{14}C method. Once the carbon is fixed in a sample of wood or the bones of an animal that dies, the clock is ‘switched on’ and the date of fixing of the carbon can be determined by the amount of ^{14}C remaining. The method is most effective for materials of ages comparable to the half-life, 5730 years, and is progressively less accurate for both younger and older samples.

The proportion of ^{14}C in atmospheric carbon (normally about 1 atom in 10^{12}) has undergone dramatic changes due to human activity in the last 100 years or so. Large scale burning of fossil fuel injects into the atmosphere carbon from which ^{14}C disappeared long ago. In the 1950s, atmospheric ^{14}C was approximately doubled by atmospheric testing of nuclear weapons. Fortunately these effects do not influence the dating of older material, but there has also been a natural fluctuation in atmospheric ^{14}C due to variations in the strength of the geomagnetic field, which partially protects the atmosphere from cosmic rays by deflecting away the primary particles (mostly protons). For precise absolute ages, a calibration of the carbon clock is required. This is, in effect, a graph of N_0 versus time. Tree rings of the very long-lived Californian bristlecone pine have served this purpose for the last part of the age range accessible to carbon dating. Calibration back to 30 000 years before the present has been achieved (Bard *et al.*, 1990), using the decay of ^{234}U to ^{230}Th (Table H.2). These are successive daughters in the decay chain of ^{238}U and are useful for dating corals and similar sedimentary materials that contain uranium, but no initial thorium.

Carbon dating has a central role in archeology and has provided a quantitative tool for the study of geological processes in the Quaternary period that are too recent to be accessible to the dating methods outlined in the following sections. However, the calibration corrections are substantial and must be applied to obtain absolute dates.

3.4 Accumulation clocks: K-Ar and U-He dating

An alternative to direct knowledge of the initial concentration, N_0, of a radioactive parent is a measurement of the concentration, D^*, of a daughter product because

$$D^* = N_0 - N = N_0(1 - e^{-\lambda t}). \tag{3.4}$$

The asterisk is used with D^* to indicate the number or concentration of radiogenic daughter nuclei produced in the time t. This is because the same isotope may occur independently of the decay and the non-radiogenic or initial component must be allowed for. Dividing Eq. (3.4) by Eq. (3.2), the unknown N_0 is eliminated,

$$\frac{D^*}{N} = \frac{1 - e^{-\lambda t}}{e^{-\lambda t}} = e^{\lambda t} - 1. \tag{3.5}$$

For a decay scheme with no initial daughter or other complications, Eq. (3.5) could be used directly in the determination of ages. This is almost true of K-Ar dating, based on the decay of a minor isotope of potassium, ^{40}K, to ^{40}Ar. The complication in this case is not a serious one. It is that only 10.5% of the ^{40}K decays yield ^{40}Ar, the remainder being β^-decays to ^{40}Ca. The ratio of the decay constant for production of ^{40}Ar, λ_{Ar}, to the total, is

$$\lambda_{\mathrm{Ar}}/\lambda = \lambda_{\mathrm{Ar}}/(\lambda_{\mathrm{Ar}} + \lambda_{\mathrm{Ca}}) = 0.105. \tag{3.6}$$

There is normally very little initial argon in igneous rocks, due to its volatility and chemical inertness. It is almost completely lost by outgassing from cooling lava. When an extrusive igneous rock solidifies, with no ^{40}Ar, its clock is set to zero. Thus for K-Ar dating the clock equation is a simple modification of Eq. (3.5),

$$^{40}\mathrm{Ar} = (\lambda_{\mathrm{Ar}}/\lambda)^{40}\mathrm{K}(\mathrm{e}^{\lambda t} - 1). \tag{3.7}$$

Estimation of the age of a rock or mineral by Eq. (3.7) requires a determination of the ratio ^{40}Ar/^{40}K. The most used method relies on independent measurements of argon and potassium. This means carefully dividing a sample into two halves that contain equal concentrations of K and Ar, and then K is measured in one half and Ar in the other. The argon measurement is made with a mass spectrometer after melting the sample in vacuum, mixing the argon released with a known quantity of isotopically separated ^{38}Ar (the ‘spike’) and removing unwanted gases. As well as allowing for the fact that mass spectrometers measure ratios very well, but not absolute quantities, by comparing three isotopes of argon this procedure provides a routine method of correcting for atmospheric contamination. Atmospheric argon has isotopic abundance ratios $^{40}\mathrm{Ar} : {}^{38}\mathrm{Ar} : {}^{36}\mathrm{Ar} = 100 : 0.063 : 0.337$. Potassium is commonly determined by a flame photometer comparison of a solution with a standard and relies on the fact that ^{40}K is a fixed fraction (0.011 67%) of total K.

An alternative method of obtaining the ratio ^{40}Ar/^{40}K in a sample is to expose it to a neutron flux in a nuclear reactor, converting a fraction of the ^{39}K present to ^{39}Ar. The ^{39}Ar thus produced is a direct measure of the potassium content, so that the Ar/K ratio can be measured by a mass spectrometer comparison of ^{40}Ar/^{39}Ar. This is more direct than separate measurements on Ar and K by different methods on separate samples. A standard sample exposed to the same neutron flux is used for calibration. The ^{40}Ar/^{39}Ar method has the advantage that step-wise heating of a solid specimen releases at different temperatures the argon held in different crystallographic sites. Then, if the sample has a history of metamorphism that has caused argon loss from less retentive sites, the ^{40}Ar/^{39}Ar ratio will be lower for the low temperature release and will date the metamorphism. This has been used to trace the evolution of the Precambrian Shield area of Canada. Correction for non-radiogenic argon is obtained from a comparison with ^{36}Ar. A graph of the ^{40}Ar/^{36}Ar ratio vs ^{39}Ar/^{36}Ar for argon released at different temperatures gives a linear plot with a gradient equal to ^{40}Ar*/^{39}Ar, where the asterisk indicates the radiogenic ^{40}Ar required for the calculation of age. A precaution is needed to avoid errors resulting from interfering nuclear reactions caused by the neutron irradiation. The presence of ^{37}Ar, which can be produced in these reactions, is an indication that this is a problem.

Argon is a tracer for the outgassing of the mantle (Section 5.2). Quenched submarine basalts from mid-ocean ridges and island hot spots, such as Loihi, off Hawaii, have ^{40}Ar/^{36}Ar ratios that are generally much higher than the atmospheric ratio. This is an indication that the primordial ^{36}Ar that accreted with the Earth is mostly in the atmosphere. We may assume either that it has always been there or that the mantle is strongly outgassed and therefore that much of the ^{40}Ar is also in the atmosphere. The discussion in Section 5.2 is consistent with the second of these alternatives.

The K-Ar method is well suited to dating igneous rocks with simple histories, especially materials that are relatively young geologically. For these materials it has the advantage that there is very little initial daughter isotope, the lower age limit of usefulness being imposed by residual argon not outgassed from a natural melt (Hayatsu and Waboso, 1985). Particular successes of the K-Ar methods are the establishment of the time scale of geomagnetic reversals (Cox *et al.*, 1963; McDougall and Tarling, 1963) and the precise dating of 1.88 million year old East African volcanic deposits (tuff) that are

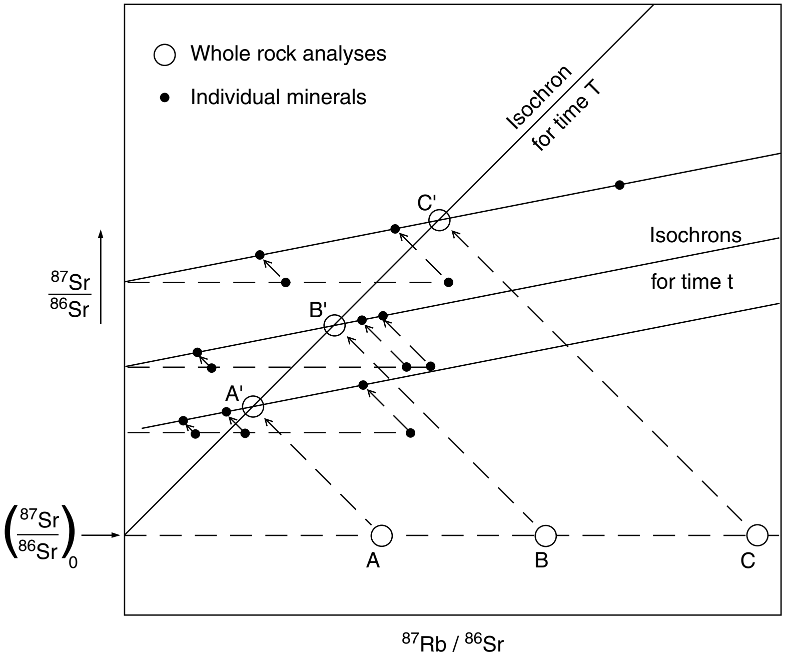


FIGURE 3.1 Rb-Sr evolution of three hypothetical rocks originating *T* years ago from a common source and undergoing simultaneous metamorphism *t* (« *T*) years ago. Original whole-rock isotopic ratios are represented by A, B, C and present ratios by A′, B′, C′. Isochrons through individual analyses for each rock date the metamorphic event and the isochron through the whole rock analyses dates the original magma differentiation.

identified with hominid remains of special interest (McDougall *et al.*, 1980; McDougall, 1981).

Being chemically inert, argon diffuses through and from minerals quite readily, so that the K-Ar method dates not the formation of a rock or mineral but the time at which it had cooled sufficiently to prevent diffusive loss of argon. This occurs at what is called the closure temperature, that is, the temperature at which a mineral grain becomes a closed system, having no further exchange with its surroundings. Different minerals have different Ar closure temperatures, according to the diffusivity of argon in them, so that a slowly cooling rock with several minerals, yielding independent K-Ar ages, may record the cooling history (e.g., McDougall and Harrison, 1999). The concept of closure temperature is not exclusive to argon. Parents and daughters of all the radioactive decay schemes diffuse at various temperatures. An idealized example, illustrated in Fig. 3.1, shows how it is possible to date both the formation of a suite of rocks and their later metamorphism. However, the problem is complicated by the fact that a closure temperature is not a sharp cut-off, but is lowered by very slow cooling.

Another accumulation clock with a gaseous daughter product is the decay of uranium and thorium, which produce ^{4}He. The He/U method was used in the first dating of rocks by Ernest Rutherford. Although He diffuses even more readily than argon, for some purposes this is an advantage. Accumulation of ^{4}He has been measured in the mineral apatite. At elevated temperatures, characteristic of depths of a few kilometres, helium diffuses from apatite as rapidly as it is produced but, as the mineral cools, the rate of helium diffusion decreases rapidly. Above about 80 °C, the helium is lost, but it is retained below about 40 °C. Measurements of He, U and Th concentrations in apatite crystals can be used to estimate how long it has been since they cooled through the range 85 °C to 40 °C. This method has been used to show that the San Gabriel Mts., in California, have been rising at about 0.3 mm/yr for last 5 million years (Blyth *et al.*, 2000).

3.5 Fission tracks

Another accumulation clock that is very simple in principle is based on the spontaneous fission of ^{238}U. This is a very rare process, occurring in only $5.4 \times 10^{-5}\%$ of ^{238}U decays, but fission fragments are very energetic and carry 40 to 50 electron charges, so that they cause very intense radiation damage along their short tracks. Each fission event produces a pair of tracks, marking the paths of the major fragments. Individual pairs of tracks are made visible for counting under a microscope by preparing polished surfaces and etching them with acid. The etchant selectively dissolves the damaged material, leaving characteristic V-shaped etch pits. Spontaneous fission of ^{235}U and ^{232}Th also occurs but is so much rarer than ^{238}U fission that it can be neglected.

Fission tracks are a radiogenic daughter for which we can be quite certain that there is no initial abundance. Thus the track count follows an accumulation clock equation analogous to Eq. (3.7),

$$T = (\lambda_{\mathrm{F}}/\lambda)\, ^{238}\mathrm{U}\left(\mathrm{e}^{\lambda t} - 1\right). \tag{3.8}$$

In this case λ_{F} is the decay constant for fission, which is very small compared with the total decay constant, λ, and T is the number of tracks caused by the available ^{238}U. The statistical problem of determining the relevant uranium abundance, corresponding to the tracks intersecting a particular plane of observation, is solved by irradiating the sample with slow neutrons in a reactor and counting the additional tracks, T_{N}, produced by neutron-induced fission of ^{235}U (no significant further fission of ^{238}U is caused). Then

$$T_{\mathrm{N}} = \phi\sigma^{235}\mathrm{U}. \tag{3.9}$$

where ϕ is the total neutron flux and $\sigma = 582 \times 10^{-28}\,\mathrm{m}^2$ is the neutron-fission cross-section of ^{235}U. The new tracks may be observed in the same plane as the original ^{238}U tracks, by re-etching, after irradiation, in which case a factor 2 arises in the comparison of Eqs. (3.8) and (3.9) because the original tracks were produced by uranium on both sides of the plane, whereas after the cut only the uranium on the remaining side can contribute. Then

$$\frac{T}{T_{\mathrm{N}}} = (2)\frac{\lambda_{\mathrm{F}}}{\lambda}\frac{^{238}\mathrm{U}}{^{235}\mathrm{U}} \cdot \frac{\left(\mathrm{e}^{\lambda t} - 1\right)}{\phi\sigma}. \tag{3.10}$$

The bracketed factor (2) does not apply if the second track count is made on a fresh plane, cut after irradiation, in which case the total count observed is $(T + T_{\mathrm{N}})$ and greater care is required to ensure that the planes compared are closely similar and the numbers of tracks statistically adequate. Although ^{235}U is a small fraction of total U, its neutron-fission cross-section, σ, is so large that there is no difficulty in making T_{N} large enough to compare with T.

Radiation damage of crystals anneals out, causing fission tracks to fade at rates that differ widely for different minerals and depend strongly on temperature. This is another example of the closure temperature problem, discussed in connection with argon. Fission track closure temperatures vary from about 120 °C (apatite) to 300 °C (sphene). Thus mild heating (to temperatures within this range) can be dated by comparing tracks in different minerals in the same rock.

Fission tracks in meteorites include a ^{235}U component due to cosmic ray-produced neutrons. The excess tracks, relative to the known meteorite ages, are thus a measure of cosmic ray exposures. This method was used to demonstrate that tektites have no observable cosmic ray exposures (Section 1.12).

3.6 The use of isochrons: Rb-Sr dating

Most of the isotopic clocks are complicated by the occurrence of initial, as well as radiogenic, daughter nuclides. This applies to the dating methods based on decays of ^{87}Rb, ^{238}U, ^{235}U and ^{147}Sm, as well as the less used ^{174}Hf, ^{176}Lu and ^{187}Re. The presence of initial abundances of daughter isotopes is not always a disadvantage. They frequently give information about the histories of the source materials of the rocks examined and may be as interesting as the ages deduced from radiogenic components. When initial daughter abundances must be allowed for in a dating scheme, a single measurement of a daughter/parent ratio does not suffice. Additional information is needed to solve for the extra unknown. In practice this means that the decay scheme must meet two conditions.

(i) A rock must have several minerals with quite different ratios of parent-to-daughter elements and it must be possible to separate them. This is normally satisfied, but it means that a rock in which only one mineral has sufficient of the relevant isotopes, or in which all the minerals have parent/daughter ratios that are too similar, cannot be dated.

(ii) A non-radiogenic (reference) isotope of the daughter element must be present for comparison.

In the case of the Rb-Sr clock, we can rewrite Eq. (3.5) with $N = {}^{87}\mathrm{Rb}$ and $D^* = {}^{87}\mathrm{Sr}^*$ and then add initial strontium, ${}^{87}\mathrm{Sr}_0$,

$$ {}^{87}\mathrm{Sr} = {}^{87}\mathrm{Sr}_0 + {}^{87}\mathrm{Sr}^* = {}^{87}\mathrm{Sr}_0 + {}^{87}\mathrm{Rb}\left(\mathrm{e}^{\lambda t} - 1\right). \tag{3.11} $$

Dividing by the abundance of the reference isotope, ${}^{86}\mathrm{Sr}$, we have the isochron (equal time) equation

$$ \frac{{}^{87}\mathrm{Sr}}{{}^{86}\mathrm{Sr}} = \left(\frac{{}^{87}\mathrm{Sr}}{{}^{86}\mathrm{Sr}}\right)_0 + \frac{{}^{87}\mathrm{Rb}}{{}^{86}\mathrm{Sr}}\left(\mathrm{e}^{\lambda t} - 1\right). \tag{3.12} $$

The initial strontium ratio, $({}^{87}\mathrm{Sr}/{}^{86}\mathrm{Sr})_0$ is the same for all minerals in an igneous rock because the two isotopes are chemically identical and, when a rock is melted, the ratio becomes uniform throughout (or very nearly so if the mass ratio is near to unity – see Section 3.9). The minerals may have quite different strontium abundances but they are homogenized with respect to isotopic ratios. Ideally, the minerals have very different ratios, Rb/Sr, of the chemically different elements, so that ${}^{87}\mathrm{Sr}/{}^{86}\mathrm{Sr}$ evolves differently with time. Thus Eq. (3.12) represents the variation with time, t, of ${}^{87}\mathrm{Sr}/{}^{86}\mathrm{Sr}$ in a suite of samples that were isotopically homogeneous at $t = 0$, such as several minerals in a single rock. If, at age t, we consider a graph of $({}^{87}\mathrm{Sr}/{}^{86}\mathrm{Sr})$ versus $({}^{87}\mathrm{Rb}/{}^{86}\mathrm{Sr})$, then it is linear, with an intercept that gives the initial Sr ratio (as would be measured in a mineral with no Rb), and a gradient $(\mathrm{e}^{\lambda t} - 1) \approx \lambda t$, which gives the age, t. The gradient is always much less than unity because ${}^{87}\mathrm{Rb}$ is long-lived compared with the Earth.

Graphical methods are helpful in visualizing the significance of isochrons, and Fig. 3.1 illustrates the use of Eq. (3.12) in a situation that allows dating of both the original emplacement and subsequent metamorphism of a suite of cogenetic rocks. Consider three rocks that are produced in a sufficiently rapid sequence to have the same age within the uncertainties of observation. Their initial whole-rock isotopic compositions are represented by the points A, B and C. They are chemically different, so that the Rb/Sr ratios cover a reasonable range, but being from a common source they are isotopically homogeneous, that is, initial strontium, $({}^{87}\mathrm{Sr}/{}^{86}\mathrm{Sr})_0$, is the same for all of them. As the rocks age, so ${}^{87}\mathrm{Rb}$ decays and for each ${}^{87}\mathrm{Rb}$ atom lost a new ${}^{87}\mathrm{Sr}$ atom is produced, causing the compositions to move along the broken lines (of gradient -1), reaching the points A′, B′ and C′ after time T. The line through these points is the T-isochron, gradient $(\mathrm{e}^{\lambda T} - 1)$. Normally we consider individual mineral analyses rather than whole rock data, but now suppose that at some time t (years ago), where $t < T$, all of the rocks represented in Fig. 3.1 were reheated sufficiently to re-homogenize the isotopes within their minerals (on a scale of centimetres), but not sufficiently to cause mixing between them (on a scale of tens or hundreds of metres). In this circumstance the whole rock analyses would be unaffected, but the mineral clocks would have been re-set to zero t years ago and so now would give isochrons with gradients corresponding to age t.

Figure 3.1 is a convenient starting point for considering also the inferences that can be drawn from initial strontium ratios. Rock C, having a high Rb/Sr ratio, accumulates radiogenic ${}^{87}\mathrm{Sr}$ faster than A or B. Thus, when its clock is reset (by the postulated reheating event), its minerals have a higher $({}^{87}\mathrm{Sr}/{}^{86}\mathrm{Sr})_0$, ratio than those in A or B, although all of the minerals in all three rocks started with the same ratio T years ago. A high initial strontium ratio characterizes a rock that was derived from a source region rich in Rb relative to Sr. The Earth's continental crust generally is such a source region. This observation is referred to in Section 5.3, in considering the evolution of the crust. Young igneous rocks with high initial strontium ratios are likely to be re-worked material with long residence times in the crust, whereas low $({}^{87}\mathrm{Sr}/{}^{86}\mathrm{Sr})_0$ (less than 0.705) probably indicates mantle-derived rocks.

Measurements of the Sr and Rb abundances are made by mass spectrometer, but a direct comparison is not possible because the two elements behave very differently when introduced to ion sources. Wet chemical methods are needed to separate rubidium from samples used for strontium analysis. Then controlled spikes of ($^{84}\mathrm{Sr} + {}^{86}\mathrm{Sr}$), or ^{87}Rb are added so that the abundances can be obtained from spectrometer measurements of ratios.

The half-life of ^{87}Rb is nearly ten times the age of the Earth. This means that variations in the $^{87}\mathrm{Sr}/^{86}\mathrm{Sr}$ ratio are quite small and high Rb/Sr ratios are required to resolve young ages. The Rb/Sr method is best suited to measurements on rocks several billion years old because neither Rb nor Sr diffuses very readily and the clock is not too trivially re-set. Moreover, the method has sufficient versatility to indicate metamorphic resetting and other complications. It is particularly useful for Precambrian geology.

3.7 U-Pb and Pb-Pb methods

Lead–uranium evolution follows equations with the same form as Eq. (3.12), with two parallel decay schemes, $^{238}\mathrm{U} \to {}^{206}\mathrm{Pb}$ and $^{235}\mathrm{U} \to {}^{207}\mathrm{Pb}$, each of which is referred to the non-radiogenic isotope ^{204}Pb,

$$\left.\begin{aligned}\frac{^{206}\mathrm{Pb}}{^{204}\mathrm{Pb}} &= \left(\frac{^{206}\mathrm{Pb}}{^{204}\mathrm{Pb}}\right)_0 + \frac{^{238}\mathrm{U}}{^{204}\mathrm{Pb}}\left(e^{\lambda_{238}t}-1\right),\\ \frac{^{207}\mathrm{Pb}}{^{204}\mathrm{Pb}} &= \left(\frac{^{207}\mathrm{Pb}}{^{204}\mathrm{Pb}}\right)_0 + \frac{^{235}\mathrm{U}}{^{204}\mathrm{Pb}}\left(e^{\lambda_{235}t}-1\right).\end{aligned}\right\} \tag{3.13}$$

Decay of uranium to lead is not immediate but proceeds in a series of steps via intermediate daughter products. The longest half-life among these products is 2.5×10^5 years for ^{234}U, great-grand-daughter of ^{238}U, and there is an isotope of the gaseous element, radon, in each of the decay series. Thus it is crucial that rocks to be dated remain closed systems, not allowing escape or introduction of any component. In some cases, especially ^{234}U, the intermediate daughters can be used as tracers, as in the study of marine sedimentation.

Equations (3.13) provide two independent clocks, but it is convenient to combine them and avoid the need to measure uranium abundances. By subtracting the initial lead from total lead in each of these equations and dividing them by one another, we obtain the radiogenic lead ratio

$$\frac{\dfrac{^{207}\mathrm{Pb}}{^{204}\mathrm{Pb}} - \left(\dfrac{^{207}\mathrm{Pb}}{^{204}\mathrm{Pb}}\right)_0}{\dfrac{^{206}\mathrm{Pb}}{^{204}\mathrm{Pb}} - \left(\dfrac{^{206}\mathrm{Pb}}{^{204}\mathrm{Pb}}\right)_0} = \frac{^{235}\mathrm{U}}{^{238}\mathrm{U}} \cdot \frac{\left(e^{\lambda_{235}t}-1\right)}{\left(e^{\lambda_{238}t}-1\right)}. \tag{3.14}$$

This ratio of radiogenic lead isotopes can be used as an indicator of the stages in the Earth's history when any particular lead sample was in contact (or lost contact) with uranium, because radiogenic ^{207}Pb increased rapidly when the shorter-lived ^{235}U was plentiful, but now ^{206}Pb is increasing faster. To make this discussion quantitative we need values for the initial lead ratios and in Section 4.3 iron meteorite lead is used for this purpose.

Equation (3.14) can be re-written in the form of an isochron equation

$$\begin{aligned}\frac{^{207}\mathrm{Pb}}{^{204}\mathrm{Pb}} = &\left[\frac{^{235}\mathrm{U}}{^{238}\mathrm{U}} \cdot \frac{\left(e^{\lambda_{235}t}-1\right)}{\left(e^{\lambda_{238}t}-1\right)}\right] \cdot \frac{^{206}\mathrm{Pb}}{^{204}\mathrm{Pb}}\\ &+ \left[\left(\frac{^{207}\mathrm{Pb}}{^{204}\mathrm{Pb}}\right)_0 - \frac{^{235}\mathrm{U}}{^{238}\mathrm{U}} \cdot \frac{\left(e^{\lambda_{235}t}-1\right)}{\left(e^{\lambda_{238}t}-1\right)}\left(\frac{^{206}\mathrm{Pb}}{^{204}\mathrm{Pb}}\right)_0\right].\end{aligned} \tag{3.15}$$

In these equations, $^{235}\mathrm{U}/^{238}\mathrm{U} = 1/137.9$ is very nearly the same in all natural materials, so that the square-bracketed terms in Eq. (3.15) are constants for a series of cogenetic samples. The first is the gradient and the second is the intercept of a lead–lead isochron, that is a graph of $^{207}\mathrm{Pb}/^{204}\mathrm{Pb}$ versus $^{206}\mathrm{Pb}/^{204}\mathrm{Pb}$, and an age can be determined from the gradient with no more information about uranium abundances than the assumption that $^{235}\mathrm{U}/^{238}\mathrm{U}$ is a known constant. Only lead isotope ratios need to be measured. In fact the constancy of the $^{235}\mathrm{U}/^{238}\mathrm{U}$ ratio is not exact, but the only known strong variation occurs in a uranium deposit at Oklo in Gabon, West Africa, parts of which are so concentrated that they operated as natural nuclear reactors two billion years ago. The uranium decay constants are the most precisely determined of any and they occur in the range most favourable for dating Precambrian events, as well as meteorites. For ^{238}U the half-life is very close to the age of the Earth and for

^{235}U the value is about 20% of this. Uranium and lead are widely distributed in the crust and in stony meteorites, and the existence of two parallel decay schemes allows a test for consistency of dating results.

If ages calculated from different decay schemes agree, they are termed concordant and assumed to be valid. The agreement may be between ^{40}Ar/^{40}K and ^{87}Sr/^{87}Rb, or between either of these and the lead isochron in Eq. (3.15), but the notion of concordance is applied particularly to Pb/U dating because the two parallel decays (Eqs. 3.13) allow a test independent of the other methods. Equation (3.15) may be re-written

$$^{206}\mathrm{Pb}/^{238}\mathrm{U} = A \times (^{207}\mathrm{Pb}/^{235}\mathrm{U}) + B, \qquad (3.16)$$

where $A = [\exp(\lambda_{238}t) - 1]/[\exp(\lambda_{235}t) - 1]$ and $B = {}^{206}\mathrm{Pb}_0/^{238}\mathrm{U} - \mathrm{A} \times {}^{207}\mathrm{Pb}_0/^{235}\mathrm{U}$.

The subscript zero refers to initial abundances, so, if there is no initial abundance of lead, $B = 0$ in Eq. (3.16). Then, since A is a universal function of age, t, concordant data fit Eq. (3.16) (with $B = 0$), a relationship termed concordia. Note that, although written here, and always plotted, as a relationship between ^{206}Pb/^{238}U and ^{207}Pb/^{235}U, concordia does not require absolute abundances of uranium isotopes. Equation (3.16), with or without B, can be multiplied through by ^{238}U and is seen to require only the ratio of uranium isotopes. But concordia is applicable only to samples with no initial lead.

Uranium, lead and intermediate daughters diffuse less readily in zircon ($ZrSiO_4$) than in other minerals, making it the most favourable mineral for lead dating. It has the further advantage that it accepts U, and also Th, as substitutes for Zr in its crystal lattice, but rejects Pb, which has a larger ionic size. Thus, zircons have very little initial Pb and a suite of cogenetic zircons will plot as a single point on concordia, if there has been no diffusion of any component. These include the inert gas, radon, which has intermediate daughter isotopes in both ^{238}U and ^{235}U decay series, so the requirement is stringent. But the real interest in concordia arose from the idea that it could be used to derive information from discordant data. A suite of zircons of age t_1, subjected to a brief heating event that re-homogenized the isotopes t_2 years ago, with no other disturbance, would lie on a chord of the concordia graph, joining the points corresponding to ages t_1 and t_2. In effect, they would be mixtures of concordant components with these ages. Unfortunately, when diffusion occurs it is found not to be so simple. If conclusions are sought from discordant lead data, then more complicated variations of isochron plots are advocated (Tera, 2003).

A particularly important early success in lead isotope measurements was the dating of the meteorites (Section 4.3). Most of the meteorites have remained unaltered and isolated from other chemical reservoirs since they were formed from an isotopically homogeneous source. They therefore give an excellent fit to a lead–lead isochron (Eq. 3.15). For terrestrial rocks, as well as meteorites, the mineral zircon is of greatest interest, not only because it has low or negligible initial lead, but because it resists diffusion of U and Pb and because it is resistant to mechanical and chemical weathering. Ion 'microprobes' that sputter very small, selected volumes from small zircon crystals allow isotopic ratios to be compared for different parts of the same crystal, and so give dates for individual zircon crystals. The oldest measured terrestrial sample is a zircon from Western Australia dated at 4.4 Ga.

3.8 ^{147}Sm-^{143}Nd and other decays

Samarium and neodymium are widely distributed, although only in trace amounts, and they are both rare-earth elements (REE). These are a sequence of chemically similar elements with a progression of properties through the periodic table that have been used as tracers of global geochemical processes. Recognition that isotopic measurements can be made precisely enough to use the very slow (10^{11} year) ^{147}Sm decay to ^{143}Nd added a new dimension to REE chemistry. ^{144}Nd is used as the reference isotope, with the proviso that measurements are invalidated if a specimen is exposed to neutrons (for example by cosmic ray bombardment) because ^{143}Nd readily absorbs neutrons to become ^{144}Nd. Although use of the Sm-Nd decay is technically difficult because of the limited range of values of ^{143}Nd/^{144}Nd and the slow decay of ^{147}Sm, it has particular advantages. Sm and Nd are less mobile even than Rb and Sr and, unlike Rb/Sr, the Sm/Nd ratio appears to have been sensibly uniform in the primitive mantle. This is evidently a consequence of the fact that these elements were not fractionated in the formation of the Earth from the solar nebula.

We can write an isochron equation, as for Rb-Sr and U-Pb,

$$\frac{^{143}\mathrm{Nd}}{^{144}\mathrm{Nd}} = \left(\frac{^{143}\mathrm{Nd}}{^{144}\mathrm{Nd}}\right)_0 + \frac{^{147}\mathrm{Sm}}{^{144}\mathrm{Nd}}\left(e^{\lambda_{\mathrm{Sm}}t} - 1\right). \quad (3.17)$$

There are three other Sm-Nd decays (Table H.1), but with half-lives that are either much longer or much shorter and they do not provide the sort of cross-checking that is available with the U-Pb clocks. Equation (3.17) represents another independent clock, but probably wider interest in the Sm-Nd method arises from the information about the source of a rock that can be obtained from the initial Nd ratio (Section 5.3). Often Nd ratios and Sr ratios are compared, or plotted against one another, to seek trends that indicate chemical evolution of the reservoirs from which they were derived.

Less used than Sm-Nd dating are methods based on the ^{176}Lu decay to ^{176}Hf, which uses ^{177}Hf as the non-radiogenic reference, and the ^{187}Re decay to ^{187}Os, with ^{186}Os as the reference. The isochron equations are exact analogues of Eqs. (3.12) and (3.17). Re and Os are siderophile elements, which means that they are found in iron meteorites and have been used to date them. Also Re and Os are strongly separated by magmatic processes, so that crustal Os is systematically different from meteoritic Os. This difference was used to demonstrate that the Os associated with Ir in the clays of the Cretaceous–Tertiary boundary is of meteoritic origin (see Section 5.5).

3.9 Isotopic fractionation

There are subtle chemical differences between the properties of molecules that have the same chemical structures, but different isotopes of the component elements. The bonding energies are the same for all isotopes, being determined by the orbital electrons, but the nuclear masses affect the molecular vibration frequencies and hence energy levels, including the zero point (low temperature) energies of the molecules. The equilibrium distribution of isotopes between interacting compounds is the result of a balance between competing effects. In the absence of thermal disturbance they would be distributed as unevenly as necessary to minimise the total energy of the system, but the randomizing effect of thermal agitation reduces the unevenness to a small bias. The resulting bias is thus a function of the temperature at which the compounds come to equilibrium and it can be used to determine that temperature. Small isotopic variations are observed for several light elements (H, C, N, O, S), permitting studies of a range of geological phenomena. The original application was to the estimation of the temperatures at which calcareous shells of marine organisms were deposited, using oxygen isotope ratios. The results obtained correlated with Pleistocene glaciations, but, as explained below, so did the isotopic composition of sea water, due to the higher concentration of light isotopes in fresh-water ice, leaving the paleotemperature estimates in doubt. Greater emphasis is now given to problems such as the evolution of sea water and conditions of magma generation and crystallization, but interest in paleotemperature data remains strong.

The equilibrium distribution of isotopes is determined quantitatively by minimizing the total free energy of a system. Although this is strictly the Gibbs free energy, G, volume and pressure changes are not involved so this is equivalent to using the Helmholtz free energy, F (see Table E.1, Appendix E),

$$F = U - TS, \quad (3.18)$$

where U is internal energy and S is entropy at temperature T. In general both U and S depend on the distribution of isotopes. If we select a convenient parameter, p, to represent this distribution then, at a particular temperature T, the condition $(\partial F/\partial p)_T = 0$ gives

$$T = (\partial U/\partial p)_T/(\partial S/\partial p)_T. \tag{3.19}$$

The principle is illustrated by considering a simple case for which U and S are calculable.

Consider an assembly of n molecules of each of two interacting compounds, AX and BX, where X is an element, common to both, that has two isotopes, X_1 and X_2. In the whole assembly there are $2n$ atoms of X, of which there is a fraction f of X_2 and $(1-f)$ of X_1. Then the distribution parameter, p, is chosen so that the numbers of molecules of the different kinds are

$$\left.\begin{aligned} &np \text{ of } AX_2,\\ &n(1-p) \text{ of } AX_1,\\ &n(2f-p) \text{ of } BX_2,\\ &n(1-2f+p) \text{ of } BX_1.\end{aligned}\right\} \tag{3.20}$$

If there is no bias in the distribution, that is X_2 is equally distributed between A and B, then $p=f$, but we are interested in the departure from this situation due to energy differences. The configurational entropy of the actual distribution is a measure of the disorder, which is greatest for $p=f$ and least if $p=0$ or $2f$, that is if X_2 attaches only to B or A. The configurational entropy, which we can represent simply by S because it is the only entropy component of interest in this context, is given by

$$S = k \ln W, \tag{3.21}$$

where k is Boltzmann's constant and W is the number of *complexions* of the system, that is the number of possible ways that X_1 and X_2 can be distributed among the $2n$ molecules in the manner of (3.20). Thus

$$W = \frac{n!}{(n-np)!(np)!}\cdot\frac{n!}{[n(2f-p)]![n(1-2f+p)]!}. \tag{3.22}$$

Since the numbers are all very large, we can use Stirling's formula to calculate logarithms of the factorials,

$$\ln N! \approx N \ln N - N, \tag{3.23}$$

to obtain

$$\begin{aligned}\ln W = &-n(1-p)\ln(1-p) - np\ln p\\ &-n(2f-p)\ln(2f-p)\\ &-n(1-2f+p)\ln(1-2f+p).\end{aligned} \tag{3.24}$$

Then by differentiating with respect to p,

$$\begin{aligned}\frac{d(\ln W)}{dp} &= n\ln\left[\frac{(1-p)(2f-p)}{p(1-2f+p)}\right]\\ &= -n\ln\left[\frac{p(1-2f+p)}{(1-p)(2f-p)}\right].\end{aligned} \tag{3.25}$$

The practical parameter used to represent the isotopic distribution is

$$\delta = \frac{(X_2/X_1) \text{ in AX}}{(X_2/X_1) \text{ standard}} - 1, \tag{3.26}$$

that is molecule AX is available for measurement. If this is a marine carbonate sample and B is sea water, which is assumed to be the same now as in the past, then B may be used as the standard for comparison and so we have

$$\delta = \frac{(X_2/X_1)_A}{(X_2/X_1)_B} - 1 = \frac{p/(1-p)}{(2f-p)/(1-2f+p)} - 1, \tag{3.27}$$

so that

$$1+\delta = \frac{p(1-2f+p)}{(2f-p)(1-p)}. \tag{3.28}$$

Comparing Eqs. (3.25) and (3.28), we see that, with $\delta \ll 1$,

$$\frac{d(\ln W)}{dp} = -n\ln(1+\delta) \approx -n\delta. \tag{3.29}$$

Thus, differentiating Eq. (3.21),

$$(\partial S/\partial p)_T = -nk\delta. \tag{3.30}$$

S would be a maximum (disorder would be maximized) if $(\partial S/\partial p)_T = 0$, that is if $\delta = 0$. However, this state is prevented by the energy differences.

We can represent the energies of the four types of molecule as

$$\left.\begin{aligned} \text{For}\quad AX_1 &: E_A,\\ AX_2 &: E_A + \Delta E_A,\\ BX_1 &: E_B,\\ BX_2 &: E_B + \Delta E_B,\end{aligned}\right\} \tag{3.31}$$

This is convenient, as it is really the differences, ΔE_A and ΔE_B, that interest us. For the isotopic distribution given by Eq. (3.20) the total energy, which we can equate to internal energy, U, because it is the only component of U that is variable in this situation, is

$$U = n[E_A + E_B + p\Delta E_A + (2f - p)\Delta E_B], \quad (3.32)$$

and therefore

$$dU/dp = n(\Delta E_A - \Delta E_B). \quad (3.33)$$

Thus, substituting Eqs. (3.30) and (3.33) in (3.19), we have

$$\delta = \frac{\Delta E_B - \Delta E_A}{kT}, \quad (3.34)$$

which says that $\delta = 0$ if $\Delta E_A = \Delta E_B$, but not otherwise.

The average energies of isotopically different molecules differ only because their energy levels are quantized. We are concerned with the vibrational energy levels that are affected by the masses of the constituent atoms. A molecule with a natural vibration frequency ν may vibrate only with energies $h\nu/2, 3h\nu/2, 5h\nu/2, \ldots$, where h is Planck's constant. The excitation occurs in multiples of $h\nu$, but there is a zero point energy, $h\nu/2$, which is the unavoidable minimum. Thus, at low temperature, at which molecules vibrate with their zero point energies, these energies vary with isotopic mass according to its effect on vibrational frequency. In general, there is a Boltzmann distribution in the occupation of the alternative energy levels, so that, at temperature T, the relative probabilities of occupation are $e^{-h\nu/2kT}, e^{-3h\nu/2kT}, e^{-5h\nu/2kT}, \ldots$. Multiplying each probability by the energy of that state and summing, we obtain the average energy

$$\overline{E} = \sum_{i=1}^{\infty} (2i - 1)\, h\nu \exp[-(2i - 1)h\nu/2kT]/2Z, \quad (3.35)$$

where

$$Z = \sum_{i=1}^{\infty} \exp[-(2i - 1)h\nu/2kT] \quad (3.36)$$

is the partition function and appears here as a normalizing factor to make the sum of all probabilities equal to unity. Equation (3.36) is a simple geometric series, giving

$$Z = e^{-h\nu/2kT} / \left[1 - e^{-h\nu/kT}\right], \quad (3.37)$$

and Eq. (3.35) can be written in terms of another standard sum (Dwight, 1961, item 33.1)

$$1 + 3x + 5x^2 + \cdots = (1 + x)/(1 - x)^2, \quad (3.38)$$

so that

$$\begin{aligned}\overline{E} &= h\nu\left(1 + e^{-h\nu/kT}\right)/2\left(1 - e^{-h\nu/kT}\right)\\ &= h\nu \coth(h\nu/2kT)/2. \quad (3.39)\end{aligned}$$

We can see why the quantization of energy levels is important to this problem by considering the classical limit of Eq. (3.39), that is $h\nu/kT \to 0$. For this situation, with the energy levels blurred into a continuum, $\overline{E} \to kT$, independently of ν. The energy is then independent of the masses of the isotopes, both ΔE_A and ΔE_B are zero and δ is zero by Eq. (3.34). At the other extreme, $h\nu/kT \to \infty$, the only state occupied is the lowest one, with zero point energy $h\nu/2$, and so $\overline{E} \to h\nu/2$. Then

$$\delta = h[(\nu_{B2} - \nu_{B1}) - (\nu_{A2} - \nu_{A1})]/2kT. \quad (3.40)$$

More generally, for arbitrary values of $h\nu/kT$ but a limited range of T, a quadratic relationship between δ and T suffices. Vibration frequencies of all modes are needed, in principle. It would be impractical/impossible to calculate them from first principles for molecules bonded to neighbours in solids and liquids but, as H. Urey first pointed out, if ν_{A1} (say) can be measured spectroscopically, then ν_{A2} or $(\nu_{A2} - \nu_{A1})$ can be calculated from it because the bond forces are the same for both isotopes and only the vibrating masses differ. The partitioning of oxygen isotopes between the shells of marine creatures and sea water has also been examined experimentally by growing them at controlled temperatures, allowing an empirical approach.

The use of oxygen isotopes in paleotemperature studies, as well as investigations of mantle-derived volcanic rocks, refers to the fractionation of ^{18}O relative to the common ^{16}O. ^{18}O is about 0.2% of common oxygen. The reference ratio is from Standard Mean Ocean Water (SMOW) and all quoted $\delta^{18}O$ values are given as departures from the SMOW ratio in parts per thousand. Thus the $\delta^{18}O$ value of a marine shell implies that it grew at a particular temperature, with the assumption that it grew in SMOW.

Oxygen isotopes fixed in carbonate with the crystal structure calcite give consistent δ values, indicating stability of the calcite, but the alternative form, aragonite, is subject to re-crystallization and is unsuitable for paleotemperature studies. A series of samples identified as calcite and laid down under apparently identical conditions by the same species allows the variation of temperature with time to be inferred. However, the isotopic composition of sea water has not been constant. The δ value of fresh water is systematically lower by about 6‰ than ocean water, and since climatic changes are associated with variable volumes of fresh water ice locked in polar regions, a bias is introduced. This can be avoided only by using two cogenetic minerals, such as calcite and calcium phosphate, with different isotopic partitioning, so that both temperature and δ^{18}O can be determined for the sea water in which shells formed.

The difference in isotopic composition between sea water and fresh water or polar ice is caused by the selective evaporation of light molecules of water, that is those with only ^{1}H and ^{16}O, leaving the remaining liquid slightly enriched in ^{2}H and ^{18}O. The lighter atoms and molecules, having higher vibrational frequencies and therefore energies, require slightly less thermal energy to escape from the binding to neighbours in the liquid. Atmospheric water has fewer of the heavy molecules than does sea water. The reverse selection of heavier isotopes for precipitation in rain and snow is less effective because the fraction of atmospheric water precipitated is much higher than the fraction of sea water evaporated. The result is an accumulation of light water in polar ice and a consequent greater enrichment of heavier isotopes in sea water during ice ages, confusing the ^{18}O/^{16}O studies of paleotemperatures.

A third isotope of oxygen, ^{17}O, has an abundance of 0.038%, slightly less than 20% of the abundance of ^{18}O, and is not normally considered in paleotemperatures but has become important in planetary and meteoritic studies. Having a mass that is half way between ^{16}O and ^{18}O, the partitioning of ^{17}O in physical and chemical processes gives δ^{17}O variations that are precisely half of the δ^{18}O variations and so give no additional information. A graph of δ^{17}O vs δ^{18}O for terrestrial oxygen samples has a gradient of ½. This is referred to as the terrestrial fractionation line. Lunar samples fall on this line but many meteorite samples, including those identified with Mars, do not. This demonstrates that the solar nebula was not isotopically homogeneous when these bodies accreted, a subject of Section 4.5.

4

Isotopic clues to the age and origin of the Solar System

4.1 Preamble

The notion that the Earth and Sun had a common origin has a long history, predating by many years modern ideas about their ages. It underlay the paradox that paralyzed geological thinking in the late 1800s: there was no known source of the Sun's energy that could warm the Earth for the apparent duration of the sedimentary record. The discovery of radioactivity by H. Becquerel in 1896 was deemed to release geological thinking from the conceptual difficulty of a very limited age for the Earth, although the release was not logically satisfying until thermonuclear fusion was recognized in the 1930s. Following the discovery of radioactivity, its two principal roles in studies of the Earth were promptly recognized. Measurements of radiogenic heat in igneous rocks, especially by Strutt (1906), and early ideas about dating, initiated by Rutherford, confirmed its significance. This chapter considers the global and Solar System questions that are illuminated by studies of isotopes; evidence for the evolution of the Earth is considered in the following chapter and radiogenic heat in Chapter 21.

Meteorites are especially important to our understanding of the early Solar System. Unlike the planets, they have suffered little modification since their common origin, 4.57×10^9 years ago. Isotopic studies on meteorites date the Solar System (Section 4.3); a precise independent age for the Earth cannot be obtained from terrestrial rocks, which have evolved in many ways from the original nebular mix. The best that can be claimed is that an average isotopic composition of crustal lead is very close to the meteorite isochron, but this leaves doubt about the validity of crustal lead as an average for the Earth as a whole.

That numerous meteorites give a good fit to a common lead isochron is evidence not only that they formed at the same time, but that lead and uranium were both isotopically homogeneous, at least in our region of the solar nebula. This is not completely true for the light elements. Selected fine grains from carbonaceous chondrites have isotopic ratios reflecting different nucleo-synthetic events (Section 4.5). They were evidently formed in atmospheres of earlier stars and maintained their integrity when they were incorporated as dust in the solar nebula. There is also a broad scale variation in the ratios of oxygen isotopes in the Solar System. This could have arisen from a dust size sorting process, driven by the Poynting–Robertson effect (Section 1.9), or by selective dissociation of gas molecules (Section 4.5) once the Sun began to radiate.

4.2 The pre-nuclear age problem

In the late 1800s geologists were seriously divided over the validity of a calculation on the cooling of the Earth that imposed a limit on its age (Burchfield, 1975). Kelvin (1863) had shown that if the Earth was cooling by diffusion of heat through the crust then the progressively thickening crust would have had a decreasing temperature gradient, reaching the present state after about 20 million years. The difficulty was that recognized sedimentary layers required hundreds of millions of years to accumulate. Considered in isolation, Kelvin's argument was a weak one. The present heat flux from the Earth (4.42×10^{13} W – Pollack *et al.*, 1993 – but less by the estimate in Kelvin's time) could be maintained for the presently understood age, 4.5×10^{9} years, with average total cooling by less than 1000 K (Problem 21.3). Internal temperatures of several thousand degrees were understood, so the limitation was not inadequacy of the heat source but the slowness of thermal diffusion in a large body. Hypotheses of convection and enhanced thermal conduction were advanced by Kelvin's contemporaries but never taken seriously because they were seen to be irrelevant. The real problem was not the Earth's heat but the energy of the Sun. Water-driven erosion and sedimentation could have occurred only as long as the surface of the Earth was warmed by the Sun. Before the discovery of nuclear reactions, there was no known mechanism to maintain the solar output for the period indicated by the sedimentary record. In retrospect it is easy to see that some new physics was needed to resolve the difficulty.

The only important source of solar energy known to pre-radioactivity physicists was gravitational collapse. As we now know, this was needed to raise the temperature of the Sun's core to the millions of degrees required to 'ignite' nuclear fusion reactions. The energy released by collapse of the Sun (mass M) to its present radius, R, and internal density structure is

$$E_{\mathrm{G}} = kGM^2/R = 6.6 \times 10^{41}\,\mathrm{J}, \tag{4.1}$$

where $k = 1.74$ is a numerical coefficient determined by the density distribution (tabulated in Problem 1.3a, Appendix J) and G is the gravitational constant. For a uniform sphere, $k = 3/5$ (Problem 1.3b) and, if this were assumed, as in the original calculation by Helmholtz (1856) and Kelvin (1862), the energy would be 2.3×10^{41} J. Equation (4.1) can be compared with the present rate of loss of energy by radiation from the Sun:

$$-\frac{\mathrm{d}E}{\mathrm{d}t} = 4\pi r_{\mathrm{E}}^2 S = 3.846 \times 10^{26}\,\mathrm{W}, \tag{4.2}$$

where r_{E} is the radius of the Earth's orbit and $S = 1370\,\mathrm{W\,m^{-2}}$ is the solar constant, the intensity of radiation at distance r_{E}. Dividing Eq. (4.1) by Eq. (4.2), we find that, at the present rate of radiation, the total gravitational energy would last $1.7 \times 10^{15}\,\mathrm{s} = 54$ million years. If we were to assume only the energy of collapse to a uniform sphere, as in the original calculation, we would obtain 19 million years. Kelvin's cooling Earth calculation derived its strength from the coincidence of his result with this value. All these estimates neglect the thermal energy stored in the Sun, which is a large fraction of the gravitational energy released and allows an even smaller age estimate.

By the end of the nineteenth century it had become clear to many geologists that the deposition of the Earth's sedimentary layers required more time than these estimates allowed. Nevertheless, the forcefulness of the physical arguments was enough to persuade some influential geologists to side with Kelvin. They included C. King, then director of the US Geological Survey who, in the conclusion to an article on the age of the Earth, published three years before the discovery of radioactivity, wrote '. . . the concordance of results between the ages of the sun and earth certainly strengthens the physical case and throws the burden of proof upon those who hold to the vaguely vast age derived from sedimentary geology.' (King, 1893).

In reviewing the evidence available to pre-radioactivity physicists, Stacey (2000) concluded that Kelvin's age-of-the-Earth paradox was inevitable. No plausible model of the Sun could provide sufficient energy to explain the sedimentary record. Even the discovery of radioactivity did not immediately solve the problem, although it suggested that a solution was possible. If the Sun were composed of 100% uranium, its radiogenic heat would be only half of the observed solar output and, in any case, the solar spectrum was incompatible with such an extreme model. The real resolution of the paradox emerged only in the 1930s with recognition of thermonuclear reactions, but such a discovery was effectively anticipated by a general rejection of the Helmholtz–Kelvin age limit. As Rutherford and Soddy (1903) wrote: 'The maintenance of solar energy, for example, no longer presents any fundamental difficulty if the internal energy of the component elements is considered to be available, i.e. if processes of sub-atomic change are going on.'

In spite of its apparent rejection, Kelvin's diffusive cooling Earth calculation had a much more prolonged influence on geological and geophysical thinking. Strutt (1906) pointed out that a 10 or 20 km layer of granite would provide enough radiogenic heat to explain the heat flux from the Earth, without involving the deep Earth at all, and suggested that radioactivity was confined to a thin, chemically distinct crust. The scene was set for a fixist view of the Earth, cooling only by thermal diffusion, which prevailed for another 60 years. Thermal models of the Earth were no more than modifications in detail of Kelvin's model until the 1960s, in spite of occasional pleas for reconsideration of convection. The final emergence, in the 1960s and 1970s of plate tectonics and a thermal history based on convective cooling was a second stage in the abandonment of Kelvin's ideas. But it is interesting to note that our use of the word *crust* developed from Kelvin's idea of a solidified layer overlying molten rock.

4.3 Meteorite isochrons and the age of the Earth

Figure 4.1 is a plot of meteorite lead isotope ratios on a lead–lead isochron (Eq. 3.15)

$$\frac{^{207}\mathrm{Pb}}{^{204}\mathrm{Pb}} = (0.613 \pm 0.014)\frac{^{206}\mathrm{Pb}}{^{204}\mathrm{Pb}} + (4.46 \pm 0.10). \tag{4.3}$$

With the decay constants in Appendix H, and taking $^{238}\mathrm{U}/^{235}\mathrm{U} = 137.88$, this gives a date of $(4.54 \pm 0.03) \times 10^9$ years for the time when isotopically homogeneous lead was isolated in various meteorite bodies with different U/Pb ratios. Iron meteorites have virtually no uranium or thorium and the least radiogenic lead found is from the troilite (iron sulphide) in the Canyon Diablo iron meteorite. This is usually regarded as the most secure point on the meteorite isochron. It is identified as *primordial* lead, that is, the lead in the original solar nebula, in Table 4.1 and used as the reference in calculating radiogenic lead ratios, as defined by Eq. (3.14) (Table 4.2).

An isochron with a precise fit, as in Eq. (4.3) and Fig. 4.1, is evidence that lead and uranium were each isotopically homogeneous in the solar nebula before the separation and accretion into solid bodies. Many of the lighter elements (C, O, N, Ne, S) were not isotopically completely homogenized and have left evidence in carbonaceous

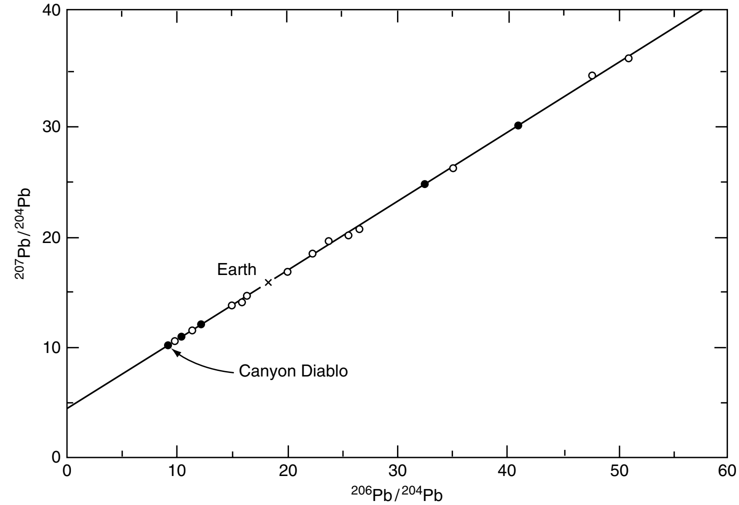


FIGURE 4.1 Lead–lead isochron for meteorites; aplot of data from two laboratories. The average terrestrial (marine sediment) data point of Chow and Patterson (1962) is shown as a cross. Some overlapping data points are omitted and values for highly radiogenic samples are well off the range plotted here, but all of the data were used to constrain the least-square fitted line (Eq. 4.3).

Table 4.1 Lead isotope ratios[a]

	^{206}Pb/^{204}Pb	^{207}Pb/^{204}Pb	^{208}Pb/^{204}Pb
Primordial (Canyon Diablo troilite)	9.307	10.294	29.476
Crustal average (marine sediments)	18.5$_8$	15.7$_7$	38.8$_7$
Ancient galena (Manitouwadge, Canada)	13.30	14.52	33.58

[a] A comparison of data by Tatsumoto *et al.* (1973) on troilite from the Canyon Diablo iron meteorite, which is widely accepted as 'primordial', that is, the least radiogenic of any natural lead, with the average crustal lead estimated by Chow and Patterson (1962) from marine sediments and an early galena, dated at 2.7 × 10^9 years and favoured by Tilton and Steiger (1965) for estimating the age of the Earth.

Table 4.2 Radiogenic lead[a]

Meteorites (Eq. 4.3)	0.613
Marine sediments (Table 4.1)	0.591
Manitouwadge galena (Table 4.1)	1.058
Easter Island[b]	0.532
Guadalupe Island[b]	0.478
East Pacific Rise[b]	0.572
Mid-Atlantic Ridge[b]	0.584

[a] Values of the ratio $\dfrac{(^{207}\mathrm{Pb}/^{204}\mathrm{Pb})_S - (^{207}\mathrm{Pb}/^{204}\mathrm{Pb})_0}{(^{206}\mathrm{Pb}/^{204}\mathrm{Pb})_S - (^{206}\mathrm{Pb}/^{204}\mathrm{Pb})_0}$ for various samples (S) compared with primordial lead (0), as in Eq. (3.14).

[b] Data from Tatsumoto (1966).

chondrites of independent nuclear sources that pre-date production of the heaviest elements in a supernova (Section 4.5). The origin of the heavy elements is considered in Section 4.4. But for the purpose of tracing the development of lead in the Earth, primordial (iron meteorite) lead is regarded as the starting point and it has been modified by additions of radiogenic lead according to subsequent contact of the lead with uranium. A requirement for the lead isotopes to fall on an isochron is that each of the uranium–lead reservoirs must have remained isolated since its formation, with no loss or addition of any component. Most of the meteorites clearly satisfy this requirement. Except for some special cases, they have remained isolated, and, apart from a brief period immediately after their formation, cool and chemically unaltered for the entire life of the Solar System.

There are no terrestrial samples comparable to the meteorites because geological activity has redistributed lead and uranium. However, an ingenious method of assessing the average terrestrial lead to find a single data point for comparison with the meteorite isochron was the analysis by Chow and Patterson (1962) of lead in ocean sediments. The idea is that erosion of continental rocks, and mixing of the products before deposition as marine sediment, produces a good approximation to the crustal average. Sediments from the different oceans gave slightly different results but, by taking a global average, Chow and Patterson obtained the best estimate that we have of average crustal lead. This gives the 'Earth' point in Fig. 4.1; the numerical values are given in Tables 4.1 and 4.2. The question of how well the marine sediment lead represents the whole Earth average is considered further in Section 5.4.

Figure 4.1 shows that the crustal average lead fits the meteorite isochron reasonably well, encouraging the conclusion that the Earth formed at the same time as the meteorites. However, the scale of the figure does not permit a close comparison and a more critical assessment is provided by the numerical values of radiogenic lead in Table 4.2. These show a wide variability in terrestrial leads with a small discrepancy between the sediment average and the meteorite isochron. We suppose that this is a real difference between the average for the crust-plus-mantle and not an artifact of inadequate sampling. There would be no discrepancy between meteorites and the Earth as a whole if the 'missing' lead could be assumed to be in the core, but this supposes that there was a delay in its separation from the mantle, that is incompatible with a more sensitive test using tungsten isotopes (Section 5.4).

The numbers in Table 4.2 give examples of *early* and *late* radiogenic lead. The lead ore at Manitouwadge, in Canada, was isolated from uranium 2.74×10^9 years ago and so represents the addition of radiogenic lead formed before that time to primordial lead. This is an example of early lead, which is relatively richer in ^{207}Pb, the decay product of the shorter-lived uranium isotope, ^{235}U. All of the other samples in the tables, including the crustal average, are seen to be biased to late lead. They must be accounted for either by a late enrichment of uranium or, equivalently, by the withdrawal of early lead. Unless crustal ore deposits, such as Manitouwadge, are vastly more extensive than we have reason to believe, they are inadequate to explain an observable early lead depletion of the whole mantle and the preferred explanation is that early lead was carried down into the core. Oversby and Ringwood (1971) found that lead is quite strongly partitioned into the iron in mixed iron–silicate melts and supposed that the lead isotope ratios could be explained in this way, but as mentioned above and in Section 5.4, this is disallowed by studies of tungsten isotopes. It is important to these arguments that lead, uranium, tungsten and hafnium (an isotope of which decays to tungsten) are heavy, neutron-rich elements that could have been produced only in a supernova. Assuming only one such event, their isotopic ratios would have been uniform in the solar nebula and not variable, as in the case of light elements from different nucleo-synthetic sources (Section 4.6). Thus lead remains an unsolved problem.

The Rb-Sr method also gives a good meteorite isochron from 'whole rock' analyses:

$$\frac{^{87}\mathrm{Sr}}{^{86}\mathrm{Sr}} = 0.0664\frac{^{87}\mathrm{Rb}}{^{86}\mathrm{Sr}} + 0.6989. \tag{4.4}$$

In this case the intercept is the value for primordial strontium obtained from achondrites, which have very low Rb contents. By Eq. (3.12) the gradient gives an age of 4.53×10^9 years. Individual meteorites have been dated also by mineral isochrons and some variability in apparent age is noted. In part this was due to a delay of perhaps a few tens of millions of years before diffusion ceased, depending on the meteorite cooling rates (Section 1.10), but it is also evident either that there is some real variation in meteorite ages or that initial strontium was not completely homogeneous. Thus, Gray *et al.* (1973) reported a value of $(^{87}\mathrm{Sr}/^{86}\mathrm{Sr})_0 = 0.698\,77$ for Rb-deficient grains in the Allende carbonaceous chondrite. Strontium with this composition must have been isolated from rubidium earlier than that in the achondrites.

The minor variability in meteorite ages and the remaining uncertainties and discrepancies are interesting details that offer clues to the accretion process and very early history of the Solar System, but cannot cast doubt on the essential conclusion that the meteorites formed from a common cloud of material about 4.57×10^9 years ago. The further conclusion that the Earth and other planets formed at the same time is hard to avoid, but is not as well documented because the Earth has had a more complicated history. But even without meteorites the age of the Earth would be constrained to a value between the oldest geological samples, about 4.4×10^9 years, and the age of the heavy elements (Section 4.4), which could possibly be extended to 6×10^9 years if no meteorite evidence were allowed.

4.4 Dating the heavy elements: orphaned decay products

That radioactive species with half-lives of 10^9 years or less, notably ^{235}U, have survived to the present time allows us to put a rough bound on their ages. An early guess by Rutherford assumed a value of 0.8 for the initial abundance ratio $(^{235}U/^{238}U)_0$. This is related to the present ratio by the difference between the two decays,

$$\left(^{235}\mathrm{U}/^{238}\mathrm{U}\right) = \left(^{235}\mathrm{U}/^{238}\mathrm{U}\right)_0 \exp[-(\lambda_{235} - \lambda_{238})\tau_0], \quad (4.5)$$

and gives $\tau_0 = 5.7 \times 10^9$ years as the time since nuclear synthesis. Certainly it is implausible that the 'age' of uranium should be close to the conventional age of the Universe inferred from the Hubble expansion, 13.7×10^9 years, which would require $(^{235}U/^{238}U)_0 = 627$. The Rutherford guess predated by many years the discovery of the neutron, which, as we now know, rapidly destroys ^{235}U by fission. The ^{235}U with which the Solar System started must have been a daughter of heavier but shorter-lived nuclides and the Rutherford guess overestimates the age of the heavy elements. This means that the interval between synthesis of these elements and their incorporation in solid bodies in the Solar System was short compared with the subsequent life of the Solar System, inviting a search for orphaned isotopes. Orphans are the decay products of short-lived isotopes that no longer exist in measurable quantities, but did so at the time of Solar System formation and have left evidence of their incorporation in meteorites by the otherwise anomalous presence of their products. Several are listed in Table H.3, but it is anomalies in the abundances of xenon isotopes that have received most attention. The subject is sometimes referred to as 'xenology'.

Xenology began in 1960 when J. H. Reynolds reported enrichment of ^{129}Xe in iodine-bearing minerals in meteorites. The only naturally occurring isotope of iodine is ^{127}I, but, according to theories of nuclear synthesis, ^{129}I was originally produced in comparable abundance. The beta-decay of ^{129}I to ^{129}Xe, with a half-life of 16.9×10^9 years, accounts for the anomalous ^{129}Xe and the short half-life imposes a tight bound on the synthesis–accretion interval. It is only the ^{129}Xe produced by decays occurring after incorporation of iodine in the meteorite minerals that appears as the excess of this isotope. This excess is a direct measure of the abundance of ^{129}I at the time of accretion.

If we make the simple assumptions that the synthesis of iodine was a very brief event (compared with the half-life of ^{129}I) and that ^{129}I and ^{127}I were produced in equal abundances, then, at the time of accretion, t years later, the ^{129}I is reduced by the factor $e^{-\lambda t}$ and it is the remaining ^{129}I that becomes the anomalous meteoritic ^{129}Xe, so that

$$^{129}\mathrm{Xe}_{\mathrm{excess}} \approx {}^{127}\mathrm{I}\, e^{-\lambda t}. \quad (4.6)$$

The more general equation describing protracted synthesis is the subject of Problem 4.2, but this is not relevant if the synthesis event was a supernova. On the basis of Eq. (4.6), Reynolds concluded that the synthesis–accretion interval was 1 to 2×10^8 years. It now appears that this is also a serious overestimate and that the interval may have been as short as 10^6 years.

Although excess ^{129}Xe is positively identified with iodine, there are relative abundance variations between the other eight stable xenon isotopes. The principal cause of this is spontaneous fission of the extinct heavy isotope of plutonium, ^{244}Pu, not to be confused with the shorter-lived ^{239}Pu that is used in the nuclear power and weapons industries. ^{244}Pu has a half-life of nearly 10^8 years (Table H.3) and 0.3% of its decays are by spontaneous fission. Its fission products include a characteristic distribution of xenon isotopes that provide a signature of its former presence. There is, of course, no stable Pu isotope that can act as a marker for the minerals that would have incorporated ^{244}Pu, in the way that ^{127}I provides a reference for the $^{129}I \rightarrow {}^{129}Xe$ decay, but chemically it appears that plutonium associated with uranium, and so the ^{244}Pu products are found in uranium-bearing minerals. ^{244}Pu fission products confirm that the synthesis–accretion interval was very short compared with the subsequent life of the Earth. The estimate cannot be made precise because accretion itself was an extended process, but also because variations in xenon isotopic abundances are still not entirely explained (Ozima and Podosek, 1999).

The production of the heavy elements considered in this section required a very intense neutron flux. Successive neutron captures occurred too rapidly to allow normal β-decay to the most stable nuclear series, along which slow nuclear build-up proceeds. The only known type of event capable of producing the required intense burst of neutrons is a supernova, so the conclusion is that a supernova occurred only 10^8 years or so (perhaps much less) before its debris was incorporated in solid bodies. Other orphans with shorter-lived parents impose much tighter limits than this for grains in carbonaceous chondrites. The implication is that the supernova provided a trigger for Solar System formation, by a mechanism considered in Section 4.6.

Another orphan that has attracted particular attention is ^{26}Mg, a product of ^{26}Al, which has a half-life of 7.2×10^5 years. The identification of isotopically anomalous Mg in meteorite samples indicates either their formation within a few million years of synthesis of Al in a supernova or prolonged intense radiation that produced ^{26}Al by spallation after the formation of solid grains. However, the spallation alternative fails to explain very detailed observations on a Ca-rich inclusion in the Allende carbonaceous chondrite (Hsu *et al.*, 2000). This inclusion had a layered structure with three crystallization events, for each of which the inferred initial $(^{26}\mathrm{Al}/^{27}\mathrm{Al})_0$ ratios were obtained. These indicated formation in three stages at intervals separated by hundreds of thousands of years, but that it occurred within about a million years of the synthesis of Al. The existence of ^{26}Al as a very early heat source in the Earth has also been suggested, but its survival long enough to be incorporated in the Earth is unlikely.

4.5 Isotopic variations of pre-Solar System origin

A supernova is needed to explain the existence of the heaviest nuclides and would certainly have contributed to the Solar System inventory of a wide range of elements. It injected this material into a cloud of both gas and dust from different nucleo-synthetic sources. Vigorous stirring of the cloud, in the wake of the supernova shock wave, mixed these ingredients, homogenizing the isotopic soup, but the mixing process was incomplete. Carbonaceous chondrites in particular contain traces of materials with isotopic ratios quite different from those in the bulk of the Solar System. They could not have survived unmixed in gaseous form and there are now observations of the microscopic solid grains that carry isotopic signatures of the pre-solar, pre-supernova period (Bernatowicz and Walker, 1997; Nittler, 2003).

Microwave spectroscopic observations of remote parts of the Galaxy have revealed gas clouds containing familiar, simple molecules, but with proportions of isotopes of common elements, such as O, S, N, as well as H, quite different from those in the Solar System. It is evident that the materials in these regions were derived from different combinations of nucleo-synthetic processes. Some of the same sort of variations are seen in the fine grains that make up the carbonaceous chondrites. One of the earliest of the isotopic 'anomalies' to be identified was ^{22}Ne, which would not have condensed in the solid dust particles of a nebula or stellar atmosphere but is explained as a decay product of ^{22}Na generated by nova outbursts. In this case the 2.6 year half-life gives little scope for a delay between its synthesis and incorporation in the SiC grains in which ^{22}Ne is found. The isotopic compositions of both Si and C in these grains are also quite different from the bulk of the Solar System. They must have formed in atmospheres of carbon-rich red giant stars with insufficient oxygen to oxidise the carbon to CO (and Si to SiO_2). Similarly, a carbon-rich environment was required for formation of nano-diamonds, with high ^{13}C, that are host to ^{15}N-rich nitrogen and xenon with isotopic abundances characteristic of slow neutron irradiation (as distinct from the neutron flash of a supernova). There were evidently at least several dying stars that contributed to the solar nebular dust with isotopic compositions characteristic of different nucleo-synthetic processes. But it appears that the trigger for Solar System formation was a supernova that produced the heaviest elements, including uranium and now extinct plutonium, as well as shorter-lived parents of the orphaned isotopes mentioned in Section 4.4 and Table H.3.

We see that the Solar System accreted from materials that preserved on a microscopic scale the isotopic signatures of several different sources. The processing of these materials in bodies

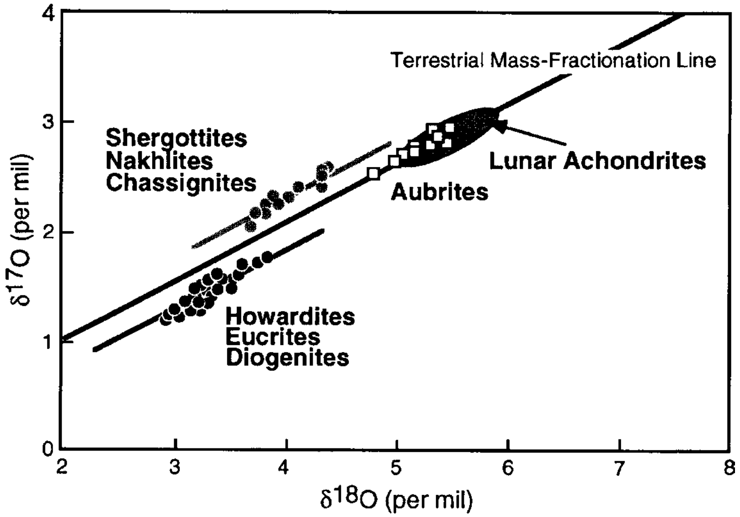


FIGURE 4.2 Oxygen mass fractionation lines for four classes of achondrite, plotted from data by R. Clayton and coworkers and reproduced, by permission, from McSween (1999).

of planetary sizes homogenized the mix, so that the isotopic variations that are apparent within them now arose only from thermally controlled physical and chemical fractionation processes, as explained in Section 3.9, and not from the preservation of pre-Solar System history. Oxygen is of particular interest in this connection because there are three isotopes, ^{16}O, ^{17}O and ^{18}O, none of which is produced radiogenically. The ratios of their abundances provide a sensitive test for isotopic homogeneity of a planet as a whole. Terrestrial samples, that is oxygen from ocean, atmosphere and rocks, all fall on a line of gradient 0.5 in a plot of δ^{17}O vs δ^{18}O, as required for an isotopically homogeneous source fractionated only by the mechanism in Section 3.9. This is known as the terrestrial mass fractionation line, but many meteorites plot well off this line and differ from one another as well as from the Earth. Figure 4.2 is a plot for several types of achondrite.

Of particular interest in Fig. 4.2 are the achondrites identified with Mars (shergottites, nakhlites, chassignites), which have their own mass fractionation line separate from that of the Earth, whereas the lunar achondrites fall on the terrestrial line. Mars accreted from material that was isotopically different from the Earth and the fractionation that has occurred within it gives a gradient 0.5 graph consistent with thermally controlled fractionation, as in Section 3.9. The inference is that there was a radial gradient in oxygen isotopic composition in the solar nebula when the planets accreted, although no single simple explanation suffices to explain all the variations between meteorites without requiring that the isotopic gradient varied with time during the accretion process, and there may well have been more than one mechanism involved.

Figure 4.3 offers another clue to a the origin of oxygen isotope variations. Refractory inclusions (CAIs) and chondrules from carbonaceous chondrites fall on a fractionation line of gradient very close to unity. Noting that all of the δ values are small, with a total range of about 4%, and that they are fractional variations for both ^{17}O/^{16}O and ^{18}O/^{16}O, a line of gradient unity passes through the origin (0,0) of a graph of total values (not δ values) of ^{17}O/^{16}O vs ^{18}O/^{16}O and therefore indicates a process that distinguishes ^{16}O from ^{17}O and ^{18}O but that ^{17}O and ^{18}O have a fixed ratio. There are plausible separation mechanisms that could operate either on molecular gases or on solid particles in the nebular cloud and perhaps both were effective. A clear choice would have important implications for the origin of the nebula.

Navon and Wasserburg (1985) and Clayton (2002) suggested that a separation process began

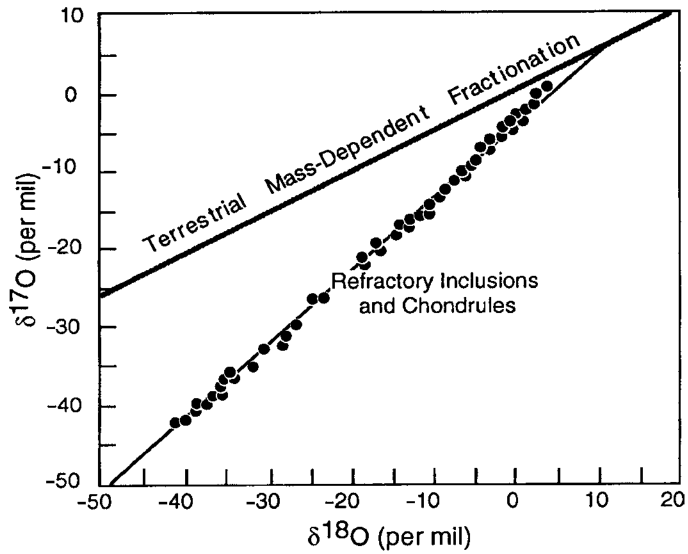


FIGURE 4.3 Oxygen isotopes in refractory inclusions (CAIs) and chondrules extracted from carbonaceous chondrites fall on a fractionation line with a gradient very close to unity, not the terrestrial mass fractionation line. Plotted from data by R. Clayton and coworkers and reproduced, by permission, from McSween (1999).

with a two-stage selective photo-dissociation of either O_2 or CO molecules in the nebular gas, exposed to strong ultra-violet radiation of the youthful Sun. A molecule is first excited to a vibrational state that is sufficiently long lived to have a precise energy level and to leave the molecule vulnerable to further UV absorption that separates it into isolated atoms. The first stage absorbs radiation of particular wavelengths that are different for molecules with different oxygen isotopes because the molecular vibration frequencies and hence energy levels are different. The more abundant $^{16}O_2$ or $^{12}C^{16}O$ molecules absorb sufficient radiation of their wavelengths in the inner part of the nebula to produce absorption lines at those wavelengths in the solar spectrum further out, so reducing the dissociation of molecules with ^{16}O, relative to ^{17}O and ^{18}O, which, being less abundant, have less effect on the spectrum. As a consequence they are more strongly dissociated than is ^{16}O. With a radial variation in the isotopic ratios of the reactive independent oxygen atoms, Clayton (2002) appealed to chemical binding to solid particles that were physically transported outwards in the nebula. This mechanism gives a separation of ^{16}O from ^{17}O and ^{18}O and so would yield materials falling on a line of unity gradient, as in Fig. 4.3.

McSween (1999, pp. 72–73) makes the alternative suggestion that the isotopes were never completely homogenized and that ^{16}O was produced virtually pure in the pre-solar supernova. The argument is that ^{17}O and ^{18}O would have been destroyed by intense radiation, because the very neutron-rich environment of the supernova that produced them was very brief compared with the following radioactivity. Then, if ^{16}O-rich grains of supernova origin were distinguishable, either by size or density, from other grains in the nebula, they would have been radially sifted by the Poynting–Robertson effect (Section 1.9), with the same result as for the molecular dissociation mechanism. Explaining all of the oxygen isotopic variations in meteorites by either mechanism requires some ingenuity but it is possible that further observations will provide the basis for a choice between them. For example, any fractionation of ^{17}O from ^{18}O would be more easily explained by the Poynting–Robertson effect because it would only require different grains from different sources, but, conversely, if IDPs are found to have very different ^{17}O/^{18}O ratios that are not seen elsewhere in the Solar System then the Poynting–Robertson effect would have separated them and evidence that it has not done so would discount this explanation.

4.6 Sequence of events in Solar System formation

Some interplanetary dust particles (IDPs) are composites of even more primitive grains that formed in the atmospheres of dying stars, but they have not been dated. Nevertheless, it is evident that grains at least similar to them collected in a cloud of gas that was probably forming for billions of years, with several inputs, before a nearby supernova provoked a gravitational collapse to form the Solar System. The subsequent development was comparatively rapid. Isotopic variations in carbonaceous chondrites demonstrate that many, perhaps most, of the original fine grains preserved their identities until well after the proto-solar nebula had condensed to a disc shape, controlled by the total angular momentum of the cloud from which it formed. Turbulence would have ensured that the nebula was very thoroughly mixed on a macroscopic scale, but did not cause grain aggregations sufficient to obscure the isotopic signatures of the several sources of nebular material. The collapse required strong interactions to damp out the random motions and produce a co-rotating disc. It is usually suggested that the supernova shock wave triggered collapse of the nebula, but it would not have reduced the turbulence. Electromagnetic damping appears to be an effective alternative. Before the supernova, the nebular cloud would have been only weakly ionized and almost non-conducting. The intensely radioactive supernova debris would have changed that, converting the cloud to a highly conducting plasma. Magnetohydrodynamic action would then have taken over and rapidly damped the turbulence. This must have happened before the Sun formed and without ionization by ultra-violet radiation.

Serious aggregation of grains began only after the increase in density of the cloud by its contraction to a disc, at which time the central concentration of mass (the Sun) would have begun to radiate. Onset of the radiation while many grains were still small caused some sorting by the Poynting–Robertson effect so that correlations between chemistry and grain size/density caused development of compositional gradients (and possibly isotopic gradients, as considered in Section 4.5). The sequence of condensation processes is inferred from abundances of ‘orphans’ of short-lived parent isotopes (Table H.3, Appendix H), especially ^{26}Mg, daughter of ^{26}Al (0.7 million year half-life). These inferences must be treated with caution because radioactivity was still strong and it included spontaneously fissioning isotopes, such as ^{244}Pu, producing neutrons that would have generated *in situ* some of the short-lived isotopes. Nevertheless, the demonstration by Hsu *et al.* (2000) of the sequential development of concentric shells of a refractory inclusion (CAI) from a carbonaceous chondrite over a period of order a million years, very shortly after the synthesis of ^{26}Al, justifies the conventional understanding that CAIs were the earliest condensed particles (that is disallowing pre-solar grains). Since they have refractory compositions it is logical to expect them to condense very early. Chondrules followed. They were still isolated droplets in the cloud when subjected to repeated transient partial (or even complete) melting and rapid cooling. This is presumed to have occurred when the Sun was at its T-Tauri stage; T-Tauri is the type example of young, very active stars, surrounded by clouds through which intense shock waves are seen to propagate, allowing the inference that chondrules were shock-heated. Significantly also, T-Tauri stars have strong magnetic fields; shock heating and cooling of chondrules in a magnetic field would have given them thermoremanent magnetizations (Section 25.2), accounting for their magnetic moments (Section 1.11).

The next stage, aggregation of millimetre-to centimetre-sized condensates into planetesimals large enough to have significant gravitational fields, is the least well understood. We may take comets as a clue to what was possible: accumulation of loose conglomerates bound by patches of volatile ices. Direct heating by the Sun to produce metallic iron is improbable but impacts between loosely compacted bodies could have done so and it appears that some meteoritic iron was formed in small bodies and did not wait for the formation of cores in large ones. Repeated fragmentation and reaggregation, as in the example in Fig. 2.2a, was normal, with repeated impact heating, but the whole process gradually dissipated the kinetic energy of turbulent or random motion, leading to progressively larger bodies. Eventual combination in a limited number of planets was probably gravitationally controlled (Section 1.2). But the meteorites give us material that has escaped the later processing, with glimpses of the early Solar System and even the sources of materials that made it.

5

Evidence of the Earth's evolutionary history

5.1 Preamble

Although we believe that the terrestrial planets all have essentially the same major chemical constituents, the atmospheres, where they exist, are very different (Tables 2.8 and 2.9). This diversity means that atmospheric composition is a sensitive indicator of the evolutionary history of the planets. The radiogenic isotope ^{40}Ar, being a decay product of ^{40}K, is of particular interest to the Earth as a whole because it would have been very rare indeed when the Earth first formed, and has leaked to the atmosphere progressively over the life of the Earth. We suppose that the potassium abundances of Venus, Earth and Mars are similar, an assumption justified for Venus by data plotted in Fig. 21.1. Then the threefold greater value of the ratio mass of atmospheric ^{40}Ar to mass of planet for the Earth than for Venus shows that the leakage of ^{40}Ar to the atmosphere has been much more effective for the Earth. This requires the convective stirring of the Earth to have been more vigorous.

We also need to explain the fact that the ratio ^{36}Ar/^{40}Ar for Venus is 270 times the terrestrial value. We can be confident that in all cases the ^{40}Ar was produced in the planets after they formed, but that ^{36}Ar was either inherited from the solar nebula or was acquired subsequently from the solar wind. Solar Ar is dominated by ^{36}Ar. But we cannot explain the very high ^{36}Ar abundance in the Venus atmosphere as an outgassing product from the planet's interior because that would require much more ^{40}Ar to have been brought up with it. It is reasonable to suppose that Venus has retained much of its primitive atmosphere, of which ^{36}Ar was a constituent. But if we assume 100 times as much as for the Earth from the ^{36}Ar/planet mass ratios, we have difficulty in explaining the fact that the N_2/planet mass ratio of Venus is also very high. There are two more pieces to be fitted into the evolutionary jig-saw puzzle: the CO_2/N_2 ratios of Venus and Mars are similar and the N_2/planet mass ratio for the Earth is 28% of that for Venus. This suggests that CO_2 and N_2 are primitive. The Earth's atmosphere has been dramatically modified by biology, which continues to remove CO_2 and makes use of N_2, although it is not clear that this leads to a net absorption of N_2 sufficient to explain its atmospheric abundance relative to Venus. But it leaves little alternative to a solar wind injection of ^{36}Ar and therefore presumably He and ^{2}H, from which the Earth was protected by the magnetosphere.

Oxygen is a special case, being a product of photosynthesis on the Earth. But oxygen is also produced by dissociation of water vapour in the upper atmosphere, where it is exposed to ultraviolet radiation, allowing the hydrogen to escape to space. The rate is limited by the coldness of the upper atmosphere, which contains little water vapour. Oxygen would have accumulated to about 20% of the present abundance if it were not continuously removed by oxidation of volcanic gases and the weathering of igneous rock. Thus it is not necessary to postulate the existence of life on Mars to explain the modest oxygen content of its atmosphere. The oxygen balance of the Earth's atmosphere is controlled by the rate at which carbon (from CO_2) is removed by photosynthesis and burial. The net rate of removal is very small compared with the carbon flux through the biosphere by growth and decay and we have no reliable independent measure of it. CO_2 is also removed by solution in the sea and incorporation in the shells of marine organisms, but that process does not affect the oxygen balance. We know also that the most voluminous material cycled through the atmosphere is water, but that appears to be an essentially balanced process that has no effect on the other constituents.

Although the Earth began with a primitive atmosphere, it is unlikely to have had much, if any, primordial crust and no continental crust. These chemically different materials must have separated from the much larger volume of the mantle and not simply been deposited last on the accreting Earth. Meteorite compositions suggest only a core and mantle and that the crust has developed over geological time, although we believe that this process started very early. Section 2.9 refers to three types of igneous crust as first, second and third stage differentiates from the mantle. Thus, crustal composition is also an indicator of the evolution of the Earth, but, in this case, we have less information from the other planets for comparison. The oceans are unique to the Earth, although that may not always have been so. It has generally been assumed that the continental crust is also unique, but it appears possible that Mars has some continental-type crust. On the other hand, it is evident that the basaltic ocean floor has equivalents on the other terrestrial planets and the Moon. But they have no sea water to act as a flux for the generation of andesitic magma from subducted ocean floor crust. Since the average lifetime of the ocean floor is about 10^8 years, only 5% of the ages of large areas of continental crust, evidence of the early history of the Earth's evolution must be sought in the continental crust (Section 5.3).

Biological activity is sensitive to its environment and the sequence of fossils that it has left in the sedimentary layers of the crust provides us with an historical record of the environment (Section 5.5). This record was the basis for a geological time scale long before the advent of nuclear dating methods, which have provided dates for the traditional geological periods (Appendix I). Each of the named periods is identified with characteristic fossils and the important feature is that they are distinct from one another. The boundaries between them mark discontinuities in the progressive evolution of life forms. We refer to mass extinctions, when many species disappeared, to be replaced by others for which environmental niches appeared or were freed up. The extinctions were consequences of environmental crises that provide clues to the evolution of the Earth itself. The ultimate causes have been contentious and, although a consensus cannot yet be claimed, the weight of evidence has shifted from an emphasis on meteorite impacts to the view that volcanic activity is the major contributor. The volcanic argument is linked to the evidence of deep convective plumes in the mantle (Chapter 12), the heat flux from the core (Chapter 18) and the power source for the geomagnetic dynamo (Chapter 21).

The evolution of the core has also been a subject of disagreement. It hangs critically on the question of a source of radiogenic heat. There is no U, Th or K in iron meteorites and most recent discussions have assumed that the radioactive content of the core is negligible. This requires the power of the geomagnetic dynamo to be derived from progressive cooling. A long-standing claim that potassium accompanied sulphur into the core (Section 2.8) is strengthened by difficulty in deriving enough heat from cooling to maintain the dynamo with a long-lived inner core (Section 21.4), but that argument can be questioned (Stacey and Loper, 2007). We know that the core has cooled much less than the mantle, leaving a thermal boundary layer at their interface. Perhaps surprisingly, the availability of radiogenic heat makes this easier to explain, because it means that the core heat flux and the dynamo can be maintained with slower cooling. However, the relatively short half life of ^{40}K introduces a complication to thermal history calculations (Chapter 23). The difficulty would be avoided by substituting uranium, but the case for that appears even weaker. It has sometimes been suggested that the formation of the core was delayed for tens of millions of years after the Earth accreted, but a simple energy argument (Section 5.4) makes that implausible. Recent comparisons of tungsten isotopic abundances in the Earth and in meteorites (summarized by Fitzgerald, 2003) confirm that formation of the core can be regarded as part of the accretion process and not a separate evolutionary episode.

5.2 Argon and helium outgassing and the Earth's potassium content

The isotope ^{40}Ar, which is produced in the Earth by decay of ^{40}K, has leaked to the atmosphere by volcanism and by weathering of crustal rocks. Being chemically inert and too heavy to escape to space, the argon is retained in the atmosphere and has accumulated to 6.55×10^{16} kg, approximately 1% of the atmosphere (Table 2.7). ^{4}He, produced by uranium and thorium decays, also leaks to the atmosphere, but is light enough to escape to space, with an average atmospheric residence time of a few times 10^5 years. Although ^{4}He is produced in the Earth in greater abundance than ^{40}Ar it is only a very minor constituent of the atmosphere. There are traces also of ^{36}Ar and ^{3}He, which are primordial, that is they were caught up in the Earth when it formed from the solar nebula and have not been supplemented by radioactive decays. However, care is required to avoid confusion with traces of these gases arriving either with the solar wind or as spallation products in interplanetary dust (see Eq. (1.7)). The very small proportions of these primordial gases, relative to the radiogenic gases ^{40}Ar and ^{4}He, demonstrate that almost all of the ^{40}Ar and ^{4}He were produced in the Earth. In the solar wind ^{36}Ar is seven times as abundant as ^{40}Ar, but in the atmosphere ^{40}Ar is 296 times as abundant.

The ratio of ^{40}Ar to ^{4}He leaking to the atmosphere indicates the K/(U + Th) ratio of the Earth. It also allows the rate of tectonic cycling of mantle material through a volcanic stage to be estimated. The abundances of K, U and Th in the crust are subject to direct measurement, but for the mantle we rely on the Ar and He trapped in rapidly quenched submarine basalts to indicate K and (U + Th) in the source regions. In principle, the primordial isotopes, ^{36}Ar and ^{3}He, give us a measure of the completeness of the outgassing process, but attempts to use them quantitatively appear to have been confused by absorption of minor gases from the quenching sea water. However, if we can obtain reliable averages of ^{40}Ar and ^{4}He, they provide a simple estimate of the K/U ratio and this gives a measure of mantle potassium, since total U in the Earth is assumed to be approximately chondritic. The degree to which the more volatile K was lost during accretion is important to the thermal budget of the Earth (Chapter 21).

Ratios of gas abundances are normally quoted by volume or, equivalently, numbers of atoms, so that for the present purpose it is convenient to convert the conventional mass ratio for solid elements (K/U) to numbers of radioactive atoms, allowing for the different atomic weights and proportions of the isotopes of the total elements:

$$\frac{^{40}\text{K atoms}}{^{238}\text{U atoms}} = 7.00 \times 10^{-4}\left(\frac{\text{K}}{\text{U}}\right)_{\text{Mass}}. \tag{5.1}$$

We also have (^{235}U atoms/^{238}U atoms) = 7.35×10^{-3} and assume a Th/U ratio that gives (^{232}Th atoms/^{238}U atoms) = 3.8. These numbers allow us to calculate from accumulation clock equations the abundance ratio ^{40}Ar/^{4}He produced by a mix with a specified K/U ratio, knowing that each ^{238}U decay produces eight ^{4}He, each ^{235}U gives seven ^{4}He and each ^{232}Th decay gives six ^{4}He, whereas on average each ^{40}K decay gives only 0.105 ^{40}Ar. For two extreme situations, current rate of production and total accumulated production over 4.5×10^9 years, we have

$$\text{Current rate: }\left(\frac{^{40}\text{Ar}}{^{4}\text{He}}\right) = 1.68 \times 10^{-5}\left(\frac{\text{K}}{\text{U}}\right)_{\text{Mass}}. \tag{5.2}$$

Total over 4.5×10^9years:

$$\left(\frac{^{40}\text{Ar}}{^{4}\text{He}}\right) = 4.53 \times 10^{-5}\left(\frac{\text{K}}{\text{U}}\right)_{\text{Mass}} \tag{5.3}$$

(Problem 5.1, Appendix J). The subscripts are a reminder that the gas ratios are expressed in terms of volumes, or numbers of atoms, but that (K/U) is the ratio by mass, facilitating direct comparison with Table 21.3.

Fisher (1975) obtained values of 10 to 20 for ^{4}He/^{40}Ar from a number of submarine basalts and concluded that the mantle K/U ratio was much lower than that of the crust. $(\mathrm{K/U})_{\mathrm{Mass}} \approx 3000$ appeared reasonable and a value as low as 1500 was not impossible. Subsequent work (e.g. Hart *et al.*, 1985) extended the available data to a larger number of samples from a wider area and also, in many cases, added observations of ^{3}He and ^{36}Ar, with the result that a more complicated picture emerged. He/Ar values vary between extremes of 0.01 and 120 and some care is required in inferring a mantle average. The low ratios are obtained from samples that are low in He, not because the argon content is high, and so do not contribute strongly to the average. It has been argued that the high ratios are biased by selective diffusion of He into melt, but after prolonged outgassing this effect would have become self-cancelling. The fact that basalts from mid-ocean ridges generally give higher ratios than basalts from island hot spots is indicative either of fractionation within the mantle or that the core contains some potassium. It is plausible that the core releases argon into the base of the mantle, where it finds its way into plumes and hence the hot spots. Acknowledging the uncertainties and an expectation that further work will throw new light on the problem, the mantle average selected here is $(^4\mathrm{He}/^{40}\mathrm{Ar}) = 12$, essentially in agreement with Fisher's (1975) original conclusion.

The relationship between ^{40}Ar/^{4}He and K/U requires a compromise between Eqs. (5.2) and (5.3). These express the fact that early radiogenic gases are richer in argon because, except for ^{235}U, ^{40}K has a shorter half-life than U or Th. Mantle degassed for the first time would give Eq. (5.3) but repeatedly degassed mantle would be nearer to Eq. (5.2). We should note also that early degassing would have been more intense. If we suppose that degassing of the Earth occurs at a rate proportional to the heat flux, then the present rate is about 60% of the average over the life of the Earth. This means that the young Earth was more strongly degassed but, of course, not much of the radiogenic gas had been produced at that time. We cannot be far wrong in adopting a coefficient of 3×10^{-5} as the compromise between Eqs. (5.2) and (5.3). With $(^4\mathrm{He}/^{40}\mathrm{Ar})_{\mathrm{Vol}} = 12$, this gives $(\mathrm{K/U})_{\mathrm{Mass}} = 2800$, which is the value adopted in Table 21.3. This is a lower value than usually quoted, and much lower than the average crustal value, but is a consequence of the argument that the mantle is strongly outgassed and that a large fraction of the argon is in the atmosphere.

Adding the crustal potassium from Table 21.3 we obtain the estimate of total terrestrial potassium, 7.14×10^{20} kg. We can compare this with the minimum amount required to produce the observed 6.5×10^{16} kg of atmospheric argon in 4.5×10^9 years. Using Eq. (3.7), and taking $^{40}\mathrm{K}/\mathrm{K}_{\mathrm{Total}} = 1.167 \times 10^{-4}$, gives $\mathrm{K}_{\mathrm{min}} = 4.7 \times 10^{20}$ kg. On this basis 66% of the argon produced in the Earth has leaked to the atmosphere. A slightly smaller fraction of the total helium may have been lost from the mantle because the leakage favoured early radiogenic gas, although He would have been lost preferentially from the crust because it diffuses more easily. Xie and Tackley (2004) modelled this problem, drawing attention to the uncertainties.

5.3 Evolution of the crust

Oceanic crust is produced at mid-ocean ridges at a rate estimated to be about 3.4 km^2/year. In global history terms it is short lived, spending about 10^8 years at the surface, and much less early in the life of the Earth, before it is subducted and re-assimilated in the mantle. The total volume generated in the life of the Earth exceeds the present volume of the crust by a factor of order 20, so no more than a tiny fraction could be converted to continental crust. But, as it drifts towards the subduction zones, the oceanic crust accumulates sediment washed off the continents and some, probably most, of this sediment is recycled into continental material. The processes of sedimentary recycling, collectively referred to by the colourful expression cannibalism, are central to the history of crustal development.

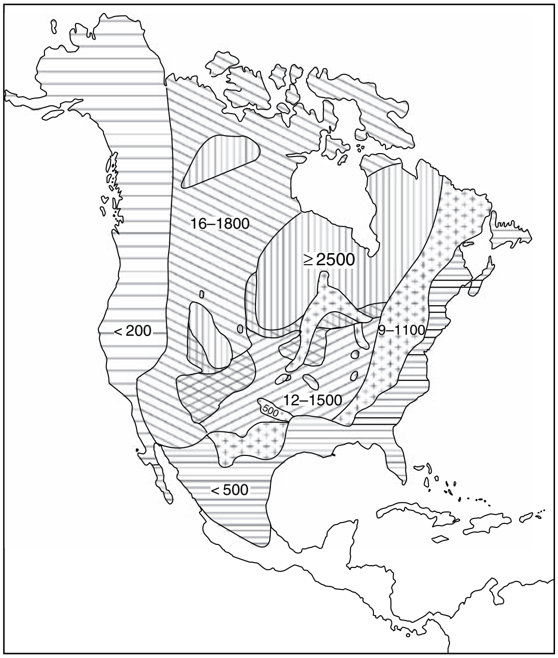


FIGURE 5.1 Age zones for basement rocks of North America. Numbers give ages in millions of years.

The rate of sediment transport to the sea is estimated by McLennan (1995) to be about 2.2×10^{13} kg/year. Approximately 20% of it reaches the ocean basins and the rest is deposited in coastal wetlands, estuaries and submarine continental margins. McLennan noted that the sediment flux increased with the advent of agriculture, so these numbers are higher than we should adopt in the present discussion. We assume a total of 10^{13} kg/year, with 2×10^{12} kg/year reaching the ocean basins. As an indication of the speed of erosion that these numbers imply, we note that the continental material above sea level, which is subject to erosion, amounts to about 3×10^{20} kg. Erosion removes this much material in about 30 million years. Nevertheless we see that the continents have ancient cores, with rocks aged up to 3.5 billion years or so and large areas with ages exceeding 2.5 billion years (Fig. 5.1). The rate of erosion is a strong function of the gradient of an eroding surface and the cratons, with modest elevation and no sharp topography, erode only very slowly. It is the areas of current or very recent tectonic activity that yield virtually all the sediment. But this sediment cannot just accumulate in the sea, or in sedimentary basins; its total volume would now greatly exceed the volume of the crust. It must be either subducted or reprocessed into continental material.

Although all of the sediment must be reworked in various ways, we should distinguish the deep ocean sediment from that accumulating on the continental margins. Much, or most, of the ocean basin sediment disappears in subduction zones, although some is scraped off on the overriding plates as accretionary wedges. There, it adds to the shallow water sediment, which is recycled into the continental crust, with a time scale of a few tens of millions of years. For much of it the fate is deep burial within the crust and eventual exhumation as consolidated sediment or metamorphic rock, but at least part of it is more thoroughly processed, perhaps by being entrained in subduction and underplating the continents, to emerge as granite, rhyolite, etc.

Sedimentary cannibalism was modelled by Veizer and Jansen (1979, 1985) as a statistical process, with the probability of any portion of crust being reworked into a more youthful form independent of time. This leads to an exponential decay with age of the volume (or area) of surviving crust. As we mention above, this statistical approach does not allow for the fact that areas of recent or current tectonic activity are generally more elevated and vulnerable to erosion, leaving the ancient cratonic areas to survive much longer than the random erosion model suggests. It appears that the exponential decay model may be applicable to the younger crustal areas, with a time constant that is only tens of millions of years, but that for old crust the time scale is much longer. The process is not well modelled by a simple exponential decay.

We mentioned that the continental crust, as well as that on the ocean floors, must ultimately have been derived from the mantle. All of the relevant processes would have been much faster early in the life of the Earth. In our discussion of thermal history (Section 23.4), we assume that the rate of crustal segregation is proportional to the convected heat flux. On this basis we can make an estimate of the fraction of fresh crustal rock that is derived directly from the mantle.

Taking the present rate of mantle heat loss to be 32×10^{12} W and the total heat loss in the life of the Earth as 12×10^{30} J, the annual fraction of this total is 8.4×10^{-11}. Assuming that the annual growth of the continental crust is the same fraction of its total production from mantle material in the life of the Earth ($\sim 7.5 \times 10^{9}$ km^3), the annual increment is about 0.6 km^3. This is quite close to the estimate by Veizer and Jansen (1985), based on chemical arguments. Thus, with 5 km^3/year of erosion, we see that about 90% of continental crust development occurs by cannibalism of earlier crust with perhaps 10% of fresh input from the mantle. The gravitational energy release by the mantle component is one of the minor items in the Earth's energy budget (Table 21.4).

Some of the processes by which the continental crust develops are probably 100% cannibalistic. Consolidated and metamorphosed sediment is unlikely to have had much contact with the mantle during its development. But, there are two tectonic processes that involve continuing chemical and isotopic exchange with the mantle. Subduction zone volcanos occur where the subducting plates have penetrated to a depth of about 100 km. The mantle heat at that depth suffices to drive off a volatile mix, which includes marine sediment, with its marker isotope, ^{10}Be, and emerges in andesite volcanos. Hot spot basalts, partial melts from material originating very deep in the mantle, almost certainly at the core–mantle boundary, are also important conveyors of mantle material to the crust.

Attempts to identify mantle components of continental crust have used isotopes of Sr (Hurley *et al.*, 1962) and Nd (DePaolo, 1981), that are produced in the decay schemes represented by Eqs. (3.11) and (3.17). The idea is that the parent elements, Rb and Sm, are more enriched in the crust, relative to the mantle, than are their daughters, so that ^{87}Sr/^{86}Sr and ^{143}Nd/^{144}Nd increase faster in the crust than in the mantle. A young igneous rock which is generated from material with a long residence time in the crust has more of the radiogenic isotope, ^{87}Sr, that is a higher initial ratio, (^{87}Sr/^{86}Sr)$_0$, than a rock derived directly from the mantle. Differentiating Eq. (3.12) and substituting for the crustal Rb/Sr ratio, the rate of increase in radiogenic strontium in the crust is

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{^{87}\mathrm{Sr}}{^{86}\mathrm{Sr}}\right) = \lambda\left(\frac{^{87}\mathrm{Rb}}{^{86}\mathrm{Sr}}\right) \approx 0.01 \times 10^{-9}\,\mathrm{year}^{-1}, \tag{5.4}$$

whereas the mantle ^{87}Sr/^{86}Sr ratio remains much closer to primordial. Thus the original date of crustal emplacement can be estimated.

These arguments are tempered by the observation of heterogeneity in the source regions of mantle-derived volcanics. ^{87}Sr/^{86}Sr for Atlantic and Pacific MORB (mid-ocean ridge basalt) is 0.7023 to 0.7027, but for OIB (ocean island or hot spot basalt) it is 0.7030 or higher. Thus the OIB source region is relatively richer in Rb. The core is quite unlikely to have influenced Rb/Sr ratios and the obvious inference is that the deep mantle has lost less Rb by differentiation into the crust. A similar difference is observed for Indian Ocean basalts, but the ratios are both higher than for corresponding Atlantic and Pacific rocks.

Particular interest attaches to the earliest development of continental crust. We argue that it was initiated by volcanism, perhaps of a rock resembling andesite. Following the earlier suggestion that this requires the presence of water at the surface, we can see a reason for a delay before it could begin to form. However, zircon is a mineral characteristic of acid, continental rocks and zircons from Western Australia have been found to have ages up to 4.4 billion years. This indicates that the delay may have been no more than 100 million years and that, within this time, the Earth had both continental-type crust and an ocean. This is the conclusion reached by Harrison *et al.* (2005 – see also Watson and Harrison, 2005) from variations in the ratios of hafnium isotopes in these zircons. The argument is essentially the same as that used to distinguish mantle-derived igneous material from reworked crust using Sr isotopes (Section 3.6). ^{176}Hf is a decay product of long-lived ^{176}Lu (Table H.1, Appendix H) and its abundance is referenced to the non-radiogenic isotope ^{177}Hf. Zircons are almost free of Lu. Thus their Hf ratios are the ambient ratios of their environments at the times of formation. The fact that the ratios are found to differ is evidence that there were already reservoirs of different ^{176}Hf/^{177}Hf ratios when the zircons formed. This requires extended periods of separation with different Lu/Hf ratios, following differentiation of the kind that produces continental crust. The first acid igneous crust, and by implication an ocean, must have appeared very early indeed.

Evidence of very early differentiation within the mantle was presented by Boyet and Carlson (2005), who pointed out that all accessible terrestrial materials have systematically higher isotopic ratios, ^{142}Nd/^{144}Nd, than do the chondritic meteorites. ^{142}Nd is a decay product of ^{146}Sm, which has a half-life of 103 million years. If, as is commonly assumed, the Earth accreted from chondritic material, with no separation of Sm from Nd, that is it has the same overall Sm/Nd ratio as the chondrites, then there must be an unsampled reservoir of material with a ratio ^{142}Nd/^{144}Nd lower than that of the chondrites. The relatively short half-life ^{146}Sm of requires that the separation of this reservoir occurred very early in the life of the Earth. This observation revives, in modified form, the idea of chemically isolated upper and lower mantles, that has generally been discarded in favour of whole mantle convection. The requirement that heat must be convectively removed from the entire lower mantle, emphasized in Section 12.6, means that the postulated isolated reservoir must be very thin. If it exists, it must be held in limited regions of D″, at the base of the mantle, that is in the crypto-continents indicated in Fig. 12.3. A simpler interpretation is that the Earth's Sm/Nd ratio does not precisely match that of the chondrites, and that no such reservoir is required. In view of the other chemical differences between the Earth and the chondrites, drawn to attention by McDonough and Sun (1995), we consider this to be more likely.

5.4 Separation of the core

A suggestion that separation of the core from the mantle may have been delayed arose first from a study of lead isotopes. Lead is moderately siderophile and would have dissolved in the core while uranium and thorium remained in the mantle. The lead in all of the terrestrial samples in Table 4.2, except for the ancient ore body, is deficient in early lead, relative to the iron meteorites. This could be explained if core formation were delayed, so that the lead dissolved in it was appreciably radiogenic but strongly biased to early lead (richer in ^{207}Pb, derived from ^{235}U), leaving the mantle biased to late lead (richer in ^{206}Pb). This hypothesis presents a serious thermo-mechanical difficulty. The gravitational energy released by settling out of the core from a homogeneous core–mantle mixture would be 1.6×10^{31} J (Stacey and Stacey, 1999), sufficient to raise the temperature of the core by more than 6300 K. The gravitational instability of a homogeneous mixture that this number represents makes its survival for tens of millions of years implausible. This is emphasized by the fact that even meteorite bodies appear to have had separated cores. As soon as any iron began to sink the heat released would trigger an avalanche and complete the process. So, we must consider that the core formed during the accretion. But the possibility of a continuing exchange between the mantle and core cannot be discounted. The stronger late lead bias of OIB lead, compared with MORB lead, is consistent with the assumption that OIB reflects a more effective gleaning of lower mantle lead into the core.

Three groups, whose work was summarized by Fitzgerald (2003), used isotopic measurements on tungsten to study this problem. ^{182}W is a decay product of ^{182}Hf, which has a half-life of 9×10^6 years, and the abundance of ^{182}W in a sample, relative to the non-radiogenic isotopes of tungsten, reflects the duration of very early contact with hafnium. The essential finding is that the ^{182}W concentration in carbonaceous chondrites is two parts in 10^4 lower than in terrestrial samples. Hafnium is a lithophile element that would have remained with the silicates during core separation, but tungsten is moderately siderophile and an appreciable fraction would have dissolved in the core. The measurements mean that mantle tungsten is slightly more radiogenic than chondritic tungsten. This indicates that the separation of the core removed some tungsten before it had acquired the chondritic complement of ^{182}W, leaving the smaller mass of mantle tungsten to accumulate all of the ^{182}W still to be released by the full complement of Hf. Interpretation in terms of the timing of core separation requires estimates of the W/Hf ratios of the mantle and chondrites and the initial abundance of ^{182}Hf, but its half-life is only comparable to the duration of the process of terrestrial accretion. The apparent slight delay in core formation is better interpreted as a delay in completion of Earth accretion after formation of the chondritic (and iron) planetesimals that plunged into it. More significant is the fact that the Earth was forming while there was still a significant amount of ^{182}Hf from the supernova that triggered Solar System formation.

5.5 The fossil record: crises and extinctions

Evolution and the proliferation of species are environmentally controlled. It is almost a case of whatever can happen will do so and therefore what happened can be interpreted simply as what was environmentally possible. The fossil record is a record of the environment. The earliest forms of life on the Earth extend back at least 3.5×10^9 years (Runnegar, 1982), but the appearance of fossilizable animals really began only 6×10^8 years ago and then expanded very rapidly. This was at, or immediately before, the beginning of the Cambrian period and appears to have coincided with a sharp increase in atmospheric oxygen. A new environmental niche appeared and evolution could rapidly occupy it. We can view the later geological periods, and the extinction events that separated them, in the same light. The paleontological record is punctuated by environmental crises. It has sometimes been postulated that the evolution of species is spasmodic, but that is a misleading interpretation. It is better described as opportunistic. Evolution can proceed rapidly with emergence of new species when conditions are favourable for them, perhaps by elimination of competitors or a changed environment. The individual geological periods are periods of environmental quiescence, more or less maintaining the status quo. The major breaks in the record occurred when environmental disturbances upset the balance.

The ultimate causes of the crises have been the subject of debate for as long as the sharp boundaries have been recognized. The debate warmed up sharply following the publication by Alvarez *et al.* (1980) of evidence of a major asteroidal or cometary impact coinciding with the boundary between the Cretaceous and Tertiary periods, 65 million years ago, when two-thirds of all species, including dinosaurs, disappeared. This boundary, and the validity of the implication that it was a direct consequence of the impact, became a focus of research on extinction events. The feature of the K-T boundary that originally attracted attention was the presence in the sedimentary record of a thin, apparently global layer rich in iridium (Ir). This is a siderophile (iron-loving) element, and its association with other siderophile elements in meteoritic proportions (Kyte *et al.*, 1985) demands an extraterrestrial source. The presence in the boundary of shocked quartz grains and tektite-like spherules is also consistent with an impact. The Ir abundance indicates that at least 10^{15} kg of chondritic material, corresponding to a body 8 km or so in diameter, was distributed around the Earth. The occurrence of such an event at or very near to the K-T boundary cannot be doubted. The questions that arise are: (1) does the impact account for the extinctions, without any other cause?, and (2) are impacts implicated in other major extinction events? Toon *et al.* (1997) reviewed the environmental effects of impacts of different sizes.

Another coincidence with the K-T boundary was the emergence of voluminous flood basalts in what is now the Deccan area of India. Between 2×10^6 km^3 and 3×10^6 km^3 of lava poured out in a period of order 10^6 years. The average rate, a few km^3 per year, is far too small to have caused a global environmental crisis, but the process was not steady. Single flows of order 10^4 km^3 in volume evidently appeared rapidly and the outgassing of such large volumes could have affected the atmosphere, and hence the climate, for several years at a time, depending on whether the sulphurous gases reached the stratosphere. Thus, extinctions caused by volcanism could be almost as sudden as those from asteroidal impact.

A feature of the K-T boundary sediments that requires a special explanation is the abundance of soot (Wolbach *et al.*, 1985). The total quantity, about 10^{15} kg, can be explained only by enormous fires. Ignition of fires that consumed more than 50% of the world's vegetation would, itself, constitute an environmental catastrophe, but it is difficult to see either an impact or volcanism as the cause and, in any case, the soot does not appear to be typical of burning vegetation. It is more plausible to suppose that a vast oil or coal deposit was exposed and burned. This could have resulted from volcanism more easily than from an impact. Soot deposition appears not to have been a single event, consistent with a global wildfire, but to have continued for a considerable time after the K-T transition itself, again indicating a fossil fuel source that could burn for a long time once ignited, or that there were multiple ignition events.

The most nearly complete extinction event occurred 250 million years ago, at the boundary between the Permian and Triassic periods, when 95% of all species disappeared. This coincided with a massive, rapid outpouring of flood basalt in what is now Siberia (Kamo *et al.*, 2003). There is no evidence of an impact, such as an iridium layer, at that time. Recognition that the two most devastating extinction events in the Phanerozoic period (the 550 million years of developed fossils) both coincided with major eruptions stimulated the search for other coincidences of extinctions and extreme volcanism. Courtillot (1999) traced the story, concluding that there are at least seven cases, including the 65 million year old K-T (Cretaceous–Tertiary) event. There is one other coincidence of an iridium layer with a transition between geological periods (Triassic–Jurassic, 200 million years ago), as reported by Olsen *et al.* (2002), but some iridium layers appear to be unrelated to extinctions. Thus, the case for a volcanic cause of mass extinctions is very strong, to some observers unassailable, although doubters remain (Wignall, 2001). The evidence for an impact cause is weaker, relying on two events for both of which the volcanic explanation appears adequate anyway, although in the case of the K-T boundary, the extinction event is identified with an impact at Chicxulub, Mexico, and the evidence for coincidence with the boundary appears convincing. The case for impacts as the causes of extinctions appears to have been overemphasized and there is a negative legacy, arising from saturation reporting of it. Courtillot (1999) even documents suppression of the volcanic alternative. The climatic effects of historical eruptions very much smaller than those responsible for the major basalt provinces are well documented (Robock, 2000, 2003; Budner and Cole-Dai, 2003), so that the effectiveness of volcanism as a cause of major extinctions cannot be doubted. Although it appears reasonable to expect similar effects from massive impacts, the present lack of evidence for an extinction event that does not also coincide with flood basalts means that confirmation of the impact effect is still lacking.

The massive basalt provinces are interpreted as surface expressions of the break-through of volcanic plumes of deep mantle origin. The debate about them focusses attention on the plume concept, first advanced by Morgan (1971) and discussed in Chapter 12. Convective plumes originating at the core–mantle boundary do not survive indefinitely, but are snuffed out and replaced by new plumes which must burn their way through the mantle with large plume heads, developing reservoirs of magma that can erupt rapidly when they eventually break through. On this basis we must suppose that mass extinctions caused by volcanism recur at irregular intervals of 50 to 100 million years.

6

Rotation, figure of the Earth and gravity

6.1 Preamble

The shape of the Earth is referred to as its figure. For some purposes it suffices to assume that the Earth is spherical. In this approximation it would interact with other astronomical bodies only by a purely central gravitational force, indistinguishable in its effect from an equal force operating on a point mass at the Earth's centre. No external torques could act on it, angular moment would be conserved, and the rotational axis would remain fixed in space even if internal motions are allowed. In this circumstance, several important geophysical effects would not occur and information about the Earth's interior derived from them would be missing. To a much better approximation the Earth is an oblate ellipsoid, quite close to the equilibrium shape resulting from a balance between the gravitational force pulling it towards a spherical shape and the centrifugal effect of rotation. The consequences of the equatorial bulge (or, equivalently, the polar flattening) are far reaching. The most important of these, from the point of view of our understanding of the Earth, is the precession, considered in the following chapter. This gives a direct measure of the moment of inertia, a crucial constraint on estimates of the internal density profile. In this chapter we consider the balance of forces causing the ellipticity and the consequent latitude variation of gravity.

How close is the Earth to the equilibrium ellipticity? A slight excess ellipticity is well documented and it is slowly decreasing. One cause is the gradual slowing of the rotation by tidal dissipation of rotational energy. From the discussion in Chapter 8 we see that, although there is a consequent gradual decrease in the equilibrium ellipticity, it is far too slow to explain the observed rate of change. More significant is a delayed recovery from the polar flattening caused by former extensive ice caps. Their retreat has left an excess ellipticity of the solid Earth, from which it is recovering on a time scale of thousands to tens of thousands of years. This is a global aspect of continuing post-glacial rebound (Chapter 9) that is most noticeable in the area around the Gulf of Bothnia (Fennoscandia) and in Eastern Canada (Laurentia). Satellite observations (Section 6.4 and Chapter 9) have given a direct measure of the rate of decrease of the ellipticity, but the process is not steady. After about 20 years of more or less regular decrease in ellipticity, an increase occurred for a few years beginning in 1988 (Cox and Chao, 2002). This is evidently a transient effect due to a mass redistribution or change in ocean circulation, but the regular decrease due to post-glacial rebound is the normal state. In principle the observations provide a measure of the viscosity of the deep mantle, but that requires details of other contributions to the excess ellipticity that are not available.

There are two important causes of apparent excess ellipticity that are unrelated to tidal friction and post-glacial rebound. One of them is relatively trivial and arises from the difference between the average tidal potential at the equator and at the poles (Chapter 8). This is not a true excess in the sense of being a departure from equilibrium but an astronomically imposed effect unrelated to the rotation of the Earth itself. More interesting is the effect of heterogeneity of the mantle. A 14 month wobble of the Earth's rotation, the Chandler wobble, discussed in Chapter 7, arises from a slight departure of the rotational axis from the axis of maximum moment of inertia. For fixed angular momentum the rotational energy is smallest for rotation about the symmetry axis and the excess energy gives a gyroscopic torque and consequent wobble that is damped with a time constant of about 30 years. The point here is that the orientation of the Earth, relative to its angular momentum axis, self-adjusts to maximize the moment of inertia about that axis. This means that the orientation is controlled by density differences in the mantle and if these are moved about by changes in the pattern of convection then they may cause true polar wander. This affects all land masses similarly and is mechanically distinct from continental drift, which is a relative motion of land masses. Making the distinction observationally is very demanding of paleomagnetic data (Chapter 22), but periodical bursts of true polar wander do appear to have occurred. So, we know that part of the excess ellipticity must be attributed to mantle heterogeneity, but with no clear indication of its importance. In Section 6.4 we suggest that it is comparable to the ellipticity in the equatorial plane, which is more than half as strong as the axial excess, and so we conclude that heterogeneity accounts for a large fraction of the observed excess.

Rotational effects arising from the interaction of the solid Earth with the atmosphere, oceans and core are discussed in Chapter 7. These are concerned with small variations on time scales of days to years but not on the time scales of mantle convection or even precession. However, discussion of the significance of rotation to the core would not be complete without mention of the rotational control of the geomagnetic field. Averaged over 10 000 years or more, the Earth's magnetic axis coincides with the rotational axis (Chapter 25), and rotation is an essential component of the complex internal motions of the core that drive the geodynamo (Chapter 24). We have no evidence that small fluctuations in the rate or axis of rotation have any effect on the field, but it appears inevitable that true polar wander, if rapid enough, would do so. This is a difficult question for paleomagnetism to answer.

6.2 Gravitational potential of a nearly spherical body

The gravitational potential, V, due to the Earth at points external to it, and in the limit on the surface itself, satisfies Laplace's equation (Eq. C.1 or C.2, Appendix C). Solutions of this equation in spherical polar coordinates, that are appropriate for the Earth's sphericity, are discussed in Appendix C. In cases of axial symmetry, the potential variation on a spherical surface can be treated as a sum of Legendre polynomials, $\mathrm{P}_l(\cos\theta)$, as defined by Eq. (C.6) and given as functions of co-latitude θ for the first few integers, l, in Table C.1. For each polynomial term the potential varies with radial distance from the coordinate origin (the centre of the Earth), r, as $r^{-(l+1)}$, as in Eqs. (C.4) and (C.7). Note that here we can ignore the r^l alternative which applies to sources of potential external to the surface considered. Thus, with complete generality, we may write the gravitational potential, $V(r,\theta)$, due to an axially symmetrical body of mass M as a sum of terms with the form of Eq. (C.7), using only the unprimed coefficients,

$$V = -\frac{GM}{r}\left(J_0P_0 - J_1\frac{a}{r}P_1(\cos\theta) - J_2\left(\frac{a}{r}\right)^2P_2(\cos\theta)\cdots\right). \quad (6.1)$$

Here a is identified as the equatorial radius, G is the gravitational constant and the coefficients J_0, J_1, $J_2, \ldots$ represent the distribution of mass. By writing the equation in this form we make these coefficients conveniently dimensionless.

Since $P_0 = 1$, we must have $J_0 = 1$, because at great distances the first term becomes dominant and this gives the potential due to a point mass (or spherically symmetrical mass). By choosing the coordinate origin to be the centre of mass, we must put $J_1 = 0$, because $P_1 = \cos\theta$ and represents an off-centre potential. Our particular interest is in the J_2 term, which is the principal one required to give the observed oblate ellipsoidal

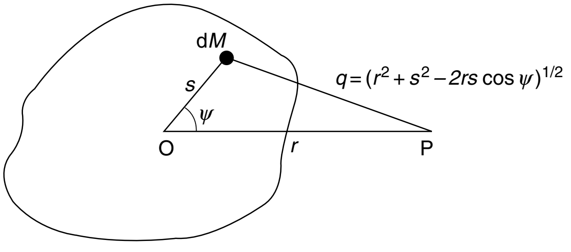


FIGURE 6.1 Geometry for the integration of gravitational potential to obtain MacCullagh's formula. The potential is calculated at a point P external to the mass M and distant r from its centre of mass, O; r is a constant in the integration, the variables being s and ψ, the coordinates of the mass element with respect to O and the line OP.

form of the geoid. All higher terms are smaller by factors of order 1000 and are neglected here, including J_4, $J_6, \ldots$ which are required for the complete representation of an ellipsoid (Eqs. C.17, C.18). Thus, with the explicit form of the function P_2, we can write the gravitational potential of the Earth as

$$V = -\frac{GM}{r} + \frac{GMa^2J_2}{r^3}\left(\frac{3}{2}\cos^2\theta - \frac{1}{2}\right). \tag{6.2}$$

Note that this gives the potential at a stationary point, not rotating with the Earth, and that a rotational potential term must be added for points rotating with the Earth. Equation (6.2) gives the potential seen by satellites, including the Moon.

It is useful to express J_2 in terms of the principal moments of inertia of the Earth. This can be done by comparing Eq. (6.2) with an alternative derivation by J. MacCullagh. Consider the geometry in Fig. 6.1. The gravitational potential at P due to the mass element dM is

$$\mathrm{d}V = -G\frac{\mathrm{d}M}{q} = -\frac{G\mathrm{d}M}{r[1 + s^2/r^2 - 2(s/r)\cos\psi]^{1/2}}. \tag{6.3}$$

This may be expanded in powers of $1/r$, ignoring terms higher than $1/r^3$, by noting that

$$\begin{aligned}\left(1 + \frac{s^2}{r^2} - 2\frac{s}{r}\cos\psi\right)^{-1/2} &\\ = 1 + \frac{s}{r}\cos\psi - \frac{1}{2}\frac{s^2}{r^2} + \frac{3}{2}\frac{s^2}{r^2}\cos^2\psi + \cdots &\\ = 1 + \frac{s}{r}\cos\psi + \frac{s^2}{r^2} - \frac{3}{2}\frac{s^2}{r^2}\sin^2\psi + \cdots. &\end{aligned} \tag{6.4}$$

Then, to this order, the total potential, obtained by integrating Eq. (6.3) with the substitution of Eq. (6.4), is a sum of four integrals, each obtained from one of the terms in Eq. (6.4):

$$\begin{aligned}V = -\frac{G}{r}\int \mathrm{d}M - \frac{G}{r^2}\int s\cos\psi\,\mathrm{d}M - \frac{G}{r^3}\int s^2\mathrm{d}M &\\ + \frac{3}{2}\frac{G}{r^3}\int s^2\sin^2\psi\,\mathrm{d}M. &\end{aligned} \tag{6.5}$$

The first integral is the potential of the centred mass, $-GM/r$. The second is zero because the centre of mass was chosen as the coordinate origin. We can transform the third term by assigning coordinates x, y, z to the elementary mass dM, so that $s^2 = (x^2 + y^2 + z^2)$ and this integral becomes

$$\begin{aligned}-\frac{G}{r^3}\int (x^2 + y^2 + z^2)\,\mathrm{d}M &= -\frac{G}{2r^3}\left[\int (y^2 + z^2)\mathrm{d}M \right.\\ &\left. + \int (x^2 + z^2)\mathrm{d}M + \int (x^2 + y^2)\mathrm{d}M\right]\\ &= -\frac{G}{2r^3}(A + B + C).\end{aligned} \tag{6.6}$$

where A, B, C are the moments of inertia of the body about the x, y, z axes. The fourth integral in Eq. (6.5) is 3/2 times the moment of inertia, I, of M about the axis OP, so that

$$V = -\frac{GM}{r} - \frac{G}{2r^3}(A + B + C - 3I)\cdots. \tag{6.7}$$

This is known as MacCullagh's formula.

To identify Eq. (6.7) more closely with Eq. (6.2) we write I in terms of A, B, C and the cosines l, m, n of the angles made by OP with the x, y, z axes:

$$I = Al^2 + Bm^2 + Cn^2, \tag{6.8}$$

where

$$l^2 + m^2 + n^2 = 1. \tag{6.9}$$

This is simplified by introducing rotational symmetry about z, so that

$$B = A. \tag{6.10}$$

Substituting for B in Eq. (6.8) and also for $(l^2 + m^2)$ by Eq. (6.9), we have

$$I = A + (C - A)n^2 \tag{6.11}$$

and Eq. (6.7) becomes

$$V = -\frac{GM}{r} - \frac{G}{2r^3}(C - A)(1 - 3n^2). \tag{6.12}$$

Since $n = \cos\theta$, this is

$$V = -\frac{GM}{r} + \frac{G}{r^3}(C-A)\left(\frac{3}{2}\cos^2\theta - \frac{1}{2}\right), \qquad (6.13)$$

which coincides with Eq. (6.2) with

$$J_2 = (C-A)/Ma^2 = 1.082\,626 \times 10^{-3}, \qquad (6.14)$$

as determined from satellite orbits (Section 9.2). This result is a step in the determination of the moment of inertia of the Earth (Section 7.2), as used in developing models of the internal variation in density (Chapter 17). In the following section a rotational term is added to Eq. (6.13) for application to points on the surface of the Earth that rotate with it.

6.3 Rotation, ellipticity and gravity

The centrifugal effect of rotation is accounted for by adding a rotational potential term to Eq. (6.13), to obtain the total geopotential at (r,θ):

$$\begin{aligned} U =& V - \frac{1}{2}\omega^2 r^2 \sin^2\theta \\ =& -\frac{GM}{r} + \frac{G}{r^3}(C-A)\left(\frac{3}{2}\cos^2\theta - \frac{1}{2}\right) \\ & -\frac{1}{2}\omega^2 r^2 \sin^2\theta, \end{aligned} \qquad (6.15)$$

where ω is the angular speed of rotation and $(r\sin\theta)$ is the distance of the surface point considered from the rotational axis. It is often convenient to write this equation in terms of latitude, ϕ, rather than co-latitude, θ,

$$\begin{aligned} U = & -\frac{GM}{r} + \frac{G}{r^3}(C-A)\left(\frac{3}{2}\sin^2\phi - \frac{1}{2}\right) \\ & -\frac{1}{2}\omega^2 r^2 \cos^2\phi. \end{aligned} \qquad (6.16)$$

The geoid is defined as the surface of constant potential, U_0, most nearly fitting the mean sea level. It has equatorial and polar radii a and c, so that the relationship between them is obtained by substituting ($r = a$, $\phi = 0$) and ($r = c$, $\phi = 90°$) in Eq. (6.16),

$$U_0 = -\frac{GM}{a} - \frac{G}{2a^3}(C-A) - \frac{1}{2}a^2\omega^2, \qquad (6.17)$$

$$U_0 = -\frac{GM}{c} + \frac{G}{c^3}(C-A), \qquad (6.18)$$

from which the flattening of the geoid is

$$f = \frac{a-c}{a} = \frac{C-A}{Ma^2}\left(\frac{a^2}{c^2} + \frac{c}{2a}\right) + \frac{1}{2}\frac{a^2 c\omega^2}{GM}. \qquad (6.19)$$

This is a first-order theory of flattening that ignores J_4 and higher contributions to the ellipsoidal form, so, to the same order in the small quantity f, the difference between a and c in the terms on the right-hand side of Eq. (6.19) can be ignored and we can write it as

$$f \approx \frac{3}{2}J_2 + \frac{1}{2}m, \qquad (6.20)$$

where J_2 is given by Eq. (6.14) and m is the ratio of the centrifugal component of gravity to total gravity at the equator.

To discuss the small departure of f from the equilibrium value for a hydrostatic Earth it is necessary to use the second-order equations relating these quantities:

$$J_2 = \frac{2}{3}f\left(1 - \frac{f}{2}\right) - \frac{m}{3}\left(1 - \frac{3}{2}m - \frac{2}{7}f\right), \qquad (6.21)$$

$$J_4 = -4f(f/5 - m/7). \qquad (6.22)$$

Numerical values of the three quantities in Eq. (6.20) are given in Table A.4 of Appendix A. It is seen that the listed geoid flattening, which is obtained using the second-order equation Eq. (6.21), is

$$f = 3.3528 \times 10^{-3}, \qquad (6.23)$$

whereas the value from Eq. (6.20) would be 3.3578×10^{-3}. The error in the first-order value is about a third of the difference between the observed and equilibrium flattening (Section 6.4).

Since the geoid is ellipsoidal, it is convenient to keep in mind the equations for an ellipsoid and the first-order approximation in terms of the flattening, f, that is used in geophysics. The conventional and readily remembered equation for an ellipse in Cartesian coordinates is

$$\frac{x^2}{a^2} + \frac{z^2}{c^2} = 1, \qquad (6.24)$$

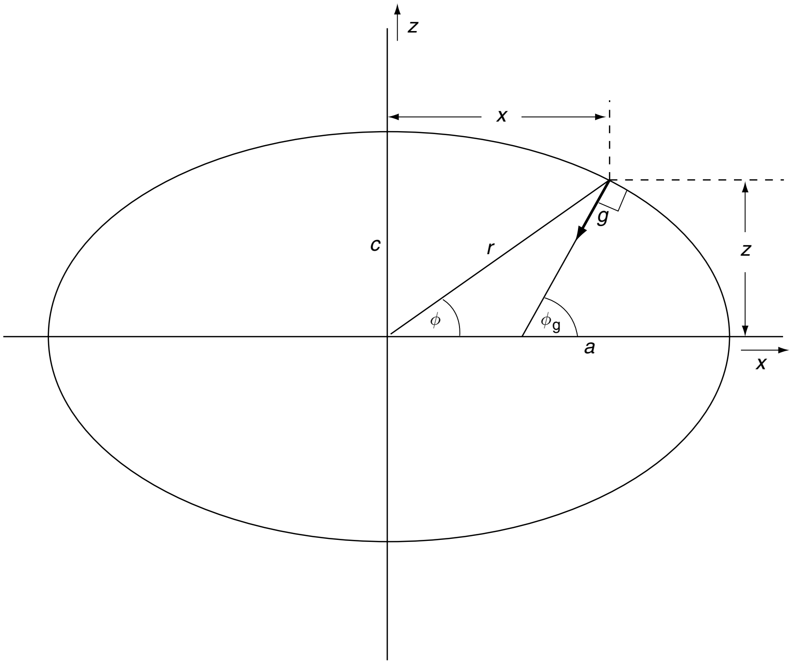


FIGURE 6.2 Oblate ellipsoidal form of the Earth with the geometrical relationship between geographic latitude, ϕ_g, and geocentric latitude, ϕ

and in terms of eccentricity, $e = (1 - c^2/a^2)^{1/2}$, or flattening, $f = (1 - c/a)$, the ellipse equation can be written in polar coordinates

$$\begin{aligned} r &= a\left[1 + \left(\frac{e^2}{1-e^2}\right)\sin^2\phi\right]^{-1/2} \\ &= a\left[1 + \left(\frac{a^2}{c^2} - 1\right)\sin^2\phi\right]^{-1/2} \\ &= a\left[1 + \frac{f(2-f)}{(1-f)^2}\sin^2\phi\right]^{-1/2}. \end{aligned} \quad (6.25)$$

The convenient first-order approximation, used to represent the geoid, is

$$r \approx a(1 - f\sin^2\phi). \quad (6.26)$$

It is necessary to note also the difference between geographic and geocentric latitudes, as illustrated in Fig. 6.2. Geographic latitude, ϕ_g, is the angle between local vertical (normal to the geoid) and the equatorial plane. Geocentric latitude, ϕ, which is used in all the equations so far, is the angle between a line to the centre of the Earth and the equatorial plane. The relationship between them is

$$\tan\phi_g = \frac{a^2}{c^2}\tan\phi = \frac{\tan\phi}{1-e^2} = \frac{\tan\phi}{(1-f)^2}. \quad (6.27)$$

A convenient and adequate approximation for translating the formula for the latitude variation of gravity from ϕ_g to ϕ (or vice versa) is

$$\sin^2\phi \approx \sin^2\phi_g - f\sin^2 2\phi_g. \quad (6.28)$$

Gravity, g, on the geoid is obtained by differentiating the geopotential, given by Eq. (6.16), with r and ϕ related by Eq. (6.26),

$$\begin{aligned} g &= -\mathrm{grad}\, U \\ &= -\left[\left(\frac{\partial U}{\partial r}\right)^2 + \left(\frac{1}{r}\frac{\partial U}{\partial \phi}\right)^2\right]^{1/2} \approx -\frac{\partial U}{\partial r}. \end{aligned} \quad (6.29)$$

The direction of g is normal to the geoid surface, as in Fig. 6.2, but the angle to the radius, $(\phi_g - \phi)$, is of order f, so to first order in small quantities the approximation in Eq. (6.29) suffices. Thus, differentiating Eq. (6.16) and substituting for $(c - a)$ by Eq. (6.14), we have

$$|g| = \frac{GM}{r^2} - \frac{3GMa^2J_2}{r^4}\left(\frac{3}{2}\sin^2\phi - \frac{1}{2}\right) - \omega^2 r\cos^2\phi. \quad (6.30)$$

The modulus is applied here because, by the definition of Eq. (6.29), g is positive upwards and so would appear as a negative quantity. We can drop the modulus by redefining g as positive downwards. The second and third terms of Eq. (6.30) are of order f times the first term, so, in substituting for r by Eq. (6.26) and expanding $\left(1 - f\sin^2\phi\right)^{-2}$ binomially, the expansion needs to be applied only to the first term. Thus

$$\begin{aligned} g &= \frac{GM}{a^2}\left(1 + 2f\sin^2\phi\right) - \frac{3GMJ_2}{a^2}\left(\frac{3}{2}\sin^2\phi - \frac{1}{2}\right) \\ &\quad - \omega^2 a\left(1 - \sin^2\phi\right) \\ &= \frac{GM}{a^2}\left[\left(1 + 2f\sin^2\phi\right) - 3J_2\left(\frac{3}{2}\sin^2\phi - \frac{1}{2}\right)\right. \\ &\quad \left. - m\left(1 - \sin^2\phi\right)\right], \end{aligned} \tag{6.31}$$

where

$$m = \omega^2 a^3/GM = 3.467\,75 \times 10^{-3} \tag{6.32}$$

is a small quantity, of order f, so that the slight difference in its definition compared with Eqs. (6.19) and (6.20) is of no consequence to this first-order theory. m is sometimes defined as the ratio of centrifugal and attractive components of equatorial gravity (instead of total gravity), in which case the value is $3.455\,76 \times 10^{-3}$.

Equatorial gravity, at $\phi = 0$, is

$$g_e = \frac{GM}{a^2}\left(1 + \frac{3}{2}J_2 - m\right) \tag{6.33}$$

and it is convenient to write g in terms of g_e. Again retaining terms only to first order in small quantities, substitution for GM/a^2 by Eq. (6.33) in Eq. (6.31) gives

$$g = g_e\left[1 + \left(2f - \frac{9}{2}J_2 + m\right)\sin^2\phi\right]. \tag{6.34}$$

By Eq. (6.20) this takes alternative forms

$$g = g_e\left[1 + \left(2m - \frac{3}{2}J_2\right)\sin^2\phi\right], \tag{6.35}$$

$$g = g_e\left[1 + \left(\frac{5}{2}m - f\right)\sin^2\phi\right]. \tag{6.36}$$

By retaining second-order terms in the theory and using Eq. (6.28) to replace ϕ by ϕ_g, we obtain the equation for the international gravity formula

$$\begin{aligned} g = g_e\left[1 + \left(\frac{5}{2}m - f - \frac{17}{14}mf\right)\sin^2\phi_g \right. \\ \left. + \left(\frac{f^2}{8} - \frac{5}{8}mf\right)\sin^2 2\phi_g\right], \end{aligned} \tag{6.37}$$

which is, with numerical values,

$$\begin{aligned} g = 9.780\,327\Big(1 + 0.005\,3024\sin^2\phi_g \\ - 0.000\,0059\sin^2 2\phi_g\Big)\ \mathrm{m\,s^{-2}}. \end{aligned} \tag{6.38}$$

This is the reference variation of gravity and departures from it are regarded as gravity anomalies.

Variations in gravity over the Earth are very small compared with the absolute value. The latitude variation is a little more than 0.5% and gravity anomalies due to internal density heterogeneities are normally very much smaller than this. The practical unit for measurement and description of gravity anomalies is the milliGal ($1\,\mathrm{mGal} \equiv 10^{-5}\,\mathrm{m\,s^{-2}}$), approximately 10^{-6} of g. Standard survey instruments nominally measure to 0.01 mGal ($10^{-8}\,g$), but difficulty in applying accurate corrections, especially those arising from terrain effects, normally raises the uncertainties to at least 1 mGal.

Gravity surveys on land are carried out on an undulating surface, not coinciding with the geoid, to which Eq. (6.38) refers, and corrections must be applied before survey data can be compared with Eq. (6.38). It is most important to know the elevations of gravity stations. The standard free air gradient or decrease in g with elevation, in areas lacking gravity anomalies or terrain problems, is $0.3086\,\mathrm{mGal\,m^{-1}}$ and a usual, but approximate, procedure is to 'correct' back to the geoid (if this is known), or to sea level, by assuming this gradient to apply. The remaining departure from the standard gravity formula is referred to as the free-air anomaly. It effectively assumes that all of the material above the geoid is collapsed down to it. An alternative presentation of gravity variations, sometimes favoured in local geological interpretation, is the Bouguer method, which assumes complete removal of the material above the geoid. For each station, this is assumed to be an infinite slab of thickness equal to the station elevation (see Problem 9.2, Appendix J).

For surface material of 'standard reference density' 2670 kg m^{-3}, the Bouguer gradient becomes 0.1967 mGal m^{-1}. Neither of these methods is satisfactory where there is more than slight topography and in general detailed topographic corrections are needed.

A comparative glance at continent-scale free-air and Bouguer anomaly maps shows that the free-air anomalies are generally much smaller. This is evidence for continental scale isostatic balance, by which surface elevation is compensated by reduced density at depth. Isostasy is considered further in Section 9.3. On a scale of 100 km or less, free-air anomalies may be much greater, because they can be supported by the strength of the lithosphere (the cool uppermost 100 km or so of the Earth).

6.4 The approach to equilibrium ellipticity

As indicated in Section 6.1, there are four causes of an excess ellipticity of the Earth, relative to hydrostatic equilibrium. It is not a simple matter to isolate them for independent study. But, although calculation of the equilibrium ellipticity itself is mathematically tricky, it is not subject to question. The complexity of the problem arises from the fact that, although the surfaces of equal density are equipotentials, their ellipticities decrease with depth. The basic reason for this can be seen by considering the equilibrium flattening of a rotating body of uniform density ρ as in Problem 6.4, which shows that flattening is proportional to ρ^{-1}. This means that in a body such as the Earth, in which density increases with depth, the ellipticities of the equipotential surfaces are affected by the material above as well as below the surfaces and so cannot be written in terms of a straightforward integral.

An approximate first order theory (e.g. Jeffreys, 1959) yields the equilibrium (hydrostatic) flattening

$$f_{\mathrm{H}} = \frac{(5/2)m}{1 + [(5/2)(1 - (3/2)C/Ma^2)]^2}, \qquad (6.39)$$

where m is given by Eq. (6.32). A higher-order treatment, with resort to numerical methods, is necessary to obtain a value that allows a satisfactory comparison with the observed flattening. Nakiboglu (1982) gives

$$f_{\mathrm{H}} = 1/299.627 = 3.337\,48 \times 10^{-3} \qquad (6.40)$$

with corresponding hydrostatic geoid coefficients

$$J_{2\mathrm{H}} = 1.072\,70_1 \times 10^{-3}, \qquad (6.41)$$

$$J_{4\mathrm{H}} = -2.992 \times 10^{-6}. \qquad (6.42)$$

Comparison of Eqs. (6.23) and (6.40) gives an excess flattening, relative to hydrostatic equilibrium, of 0.5%, which corresponds to a difference in equatorial and polar radii 100 m greater than equilibrium. We can note that if equilibrium is assumed then Eq. (6.39) allows an estimate of the moment of inertia from the surface flattening. On this basis the observed flattening of the Earth gives $C/Ma^2 = 0.3309$, which is quite close to the measured value, 0.3307. Thus, it is reasonable to apply Eq. (6.39) to Mars, for which observations of precession give a value of the moment of inertia, to show that, within the accuracy of observations, it has an equilibrium ellipticity.

The dynamic oblateness, J_2, (Eq. 6.14), is slowly decreasing as the Earth adjusts to diminished polar ice caps. Cox and Chao (2002) have documented this effect, pointing out that for two decades of precise measurement, 1979–1998, there was a more or less steady rate of change, $\mathrm{d}J_2/\mathrm{d}t \approx -2.8 \times 10^{-11}$ per year, but that for four years, 1998–2002, the trend was temporarily reversed. Chao *et al.* (2003) suggested that this could be explained by a climatically driven decade-scale oscillation in the Pacific Ocean. The earlier trend has resumed and is interpreted as the long-term effect attributable to post-glacial rebound, a topic of Section 9.5.

The total excess J_2, $\Delta J_2 = 9.9 \times 10^{-6}$, the difference between Eqs. (6.14) and (6.41), suggests a very long relaxation time, $\Delta J_2/(\mathrm{d}J_2/\mathrm{d}t) = 350\,000$ years, but this is meaningless. We need to consider more carefully what ΔJ_2 means. As mentioned in Section 6.1, there are four causes of the excess ellipticity and not all of them can be considered as departures from equilibrium. The tidal effect, that is the ellipticity resulting from the difference between the tidal potential at the poles and the average on the equator (Chapter 8) must obviously be subtracted, but only reduces the effective ΔJ_2 by 1.0×10^{-8}. The contribution to $\mathrm{d}J_2/\mathrm{d}t$ by slowing rotation, caused by tidal friction (Chapter 8) is seen to be negligible by noting that $\mathrm{d}\ln\omega/\mathrm{d}t = -2.8 \times 10^{-10}\,\mathrm{year}^{-1}$ and that the equilibrium J_2 varies with rotational speed ω as ω^2, so that $(\mathrm{d}\ln J_2/\mathrm{d}t)_{\mathrm{tidal}} = -5.6 \times 10^{-10}$ year^{-1} and therefore $(\mathrm{d}J_2/\mathrm{d}t)_{\mathrm{tidal}} = -6 \times 10^{-13}$ year^{-1}. For any plausible lag of the flattening this corresponds to a negligible contribution to ΔJ_2. In fact it is smaller than the human contribution to $\mathrm{d}J_2/\mathrm{d}t$ caused by impoundment of water in reservoirs, for which Chao (1995) gives a minimum estimate of $-1 \times 10^{-12}\,\mathrm{year}^{-1}$.

The major problem is heterogeneity of the mantle and the fact that the Earth self-adjusts its orientation to minimize rotational energy by aligning the heterogeneities to maximize the moment of inertia about the rotational axis. We have no reliable way of estimating the resulting contribution to ΔJ_2, but an idea of how big it is likely to be is obtained from the ellipticity in the equatorial plane. From the spherical harmonic coefficients of the gravitational potential in Table 9.1 we see that the coefficients representing equatorial ellipticity give $[(C_2{}^2)^2 + (S_2{}^2)^2] = 2.8 \times 10^{-6}$, compared with the excess axial ellipticity, 4.4×10^{-6}, obtained from the difference between observed and equilibrium values of $C_2{}^0$. Thus, subtracting the equilibrium ellipticity, we see the Earth as a triaxial ellipsoid with moments of inertia $C > B > A$ such that $(C - B)$ and $(B - A)$ are comparable. This is what would be expected statistically, so we have no basis for supposing that more than a small fraction of the excess ellipticity is attributable to glacial depression and we have no direct way of estimating the magnitude of this fraction. Note that Eqs. (6.11)to (6.18) assume Eq. (6.10) and therefore that the value of A is really $(A + B)/2$.

The observed progressive decrease in J_2 gives a tantalizing glimpse of deep mantle rheology. If the component of ΔJ_2 representing residual glacial depression of the poles were precisely known, then its decay, together with the other low order geoid harmonics (Chapter 9), would give a measure of the depth variation of mantle viscosity. Mitrovica and Peltier (1993) and Han and Wahr (1995) discussed this problem, concluding that available data did not suffice for a clear conclusion, and it remains true that there is no satisfactory independent evidence of ΔJ_2. However, post-glacial rebound does provide estimates of mantle rheology and most recent studies agree that there is an increase in viscosity with depth in the mantle by a factor of order 100 (e.g. Mitrovica and Forte, 1997; Kaufman and Lambeck, 2000).

Variations in J_2 and in the rotation rate are linked. In considering the effect of the tidal slowing, we can apply Eq. (6.20) to the hydrostatic state of the Earth and see that both f and J_2 vary with rotation in the manner of m, which depends on ω^2. Thus the equilibrium value of J_2, that is $J_{2\mathrm{H}}$, is proportional to ω^2 and so the variation caused by the slowing rotation is

$$\frac{1}{J_{2\mathrm{H}}}\left(\frac{\mathrm{d}J_{2\mathrm{H}}}{\mathrm{d}t}\right) = \frac{2}{\omega}\frac{\mathrm{d}\omega}{\mathrm{d}t}. \tag{6.43}$$

With the total slowing by lunar and solar tides given by Eq. (8.31), $\omega = -6.5 \times 10^{-22}\,\mathrm{rad\,s}^{-2}$, this yields

$$\begin{aligned} \mathrm{d}J_{2\mathrm{H}}/\mathrm{d}t &= -\,1.9 \times 10^{-20}\,\mathrm{s}^{-1} \\ &= -\,6.1 \times 10^{-13}\,\mathrm{year}^{-1}. \end{aligned} \tag{6.44}$$

This is below the level of detectability. If we can estimate the time constant, τ, for relaxation of the excess ellipticity, then the excess attributable to the slowing rotation is

$$\delta J_{2\omega} = (\mathrm{d}J_{2\mathrm{H}}/\mathrm{d}t)\tau. \tag{6.45}$$

For $\tau \approx 10^4$ years, $\delta J_{2\omega} \approx 6.1 \times 10^{-9}$, which is close to the resolution limit for measurements of J_2 and is not a significant effect.

Now consider the change in rotation rate due to the variation of J_2 caused by the post-glacial rebound (Section 9.6). The angular momentum of the Earth is conserved in this process, so that

$$\mathrm{d}(C\omega)/\mathrm{d}t = 0, \tag{6.46}$$

where C is the axial moment of inertia, and therefore

$$\dot{\omega}/\omega = -\dot{C}/C. \tag{6.47}$$

In relating this to $\dot{J}_{2R}$, the rebound component of dJ_2/dt, it is convenient to introduce the parameter $(C/Ma^2) = 0.330\,695$, because this is determined by the Earth's density profile and is not a function of time (i.e. $C \propto a^2$). Then, rewriting the definition of J_2 (Eq. 6.14),

$$J_2 = \frac{C - A}{Ma^2} = \left(1 - \frac{A}{C}\right)\left(\frac{C}{Ma^2}\right), \tag{6.48}$$

and noting that $\dot{A} = -\dot{C}/2$, we have

$$\dot{J}_{2R} = \frac{\dot{C}}{C}\left(\frac{1}{2} + \frac{A}{C}\right)\left(\frac{C}{Ma^2}\right) = 0.4948\frac{\dot{C}}{C}. \tag{6.49}$$

Thus, by Eq. (6.47), there is a rebound component of rotational acceleration

$$\dot{\omega} = -\omega\dot{C}/C = -2.02\omega\dot{J}_{2R}. \tag{6.50}$$

Identifying $\dot{J}_2 = -2.8 \times 10^{-11}\,\text{year}^{-1}$ with $\dot{J}_{2R}$ we have a rebound component of rotational acceleration

$$\dot{\omega}_R/\omega = 5.7 \times 10^{-11}\,\text{year}^{-1} = 1.8 \times 10^{-18}\,\text{s}^{-1}, \tag{6.51}$$

and therefore

$$\dot{\omega}_R = 1.3 \times 10^{-22}\,\text{rad s}^{-2}. \tag{6.52}$$

This opposes the frictional slowing by Eq. (8.31), $\dot{\omega}_{\text{tidal}} = -6.5 \times 10^{-22}\,\text{rad s}^{-2}$, giving a net slowing $\dot{\omega} = -5.2 \times 10^{-22}\,\text{rad s}^{-2}$.

In addition to the tidal and rebound contributions to $\dot{\omega}$ there are irregular effects of angular momentum exchange with the atmosphere, oceans and core, but if we consider an average over 1000 years or more we have some prospect that these short-term effects are averaged out. This is the approach of Stephenson and Morrison (1995), who concluded, from 2000 years of eclipse data, that allowance for the rebound effect removed what would otherwise be a discrepancy in tidal friction observations. However, the rotational effects of glaciation and rebound occur over thousands to tens of thousands of years and are brief transients on the time scale of tidal friction which acts over the entire life of the Earth (Sections 8.3, 8.4).

7

Precession, wobble and rotational irregularities

7.1 Preamble

The axis of the Earth's rotation is inclined to the pole of the ecliptic (normal to the orbital plane) by $23^{\circ}.45$. This gives us the seasons. But the axis does not maintain a constant orientation in space. Gravitational interactions between the equatorial bulge and the Moon and Sun cause a slow precession of the axis about the ecliptic pole. The axis describes a cone with a 47° angle in a period of 25 730 years, fast enough for precise astronomical measurement, but not so fast that navigation by the stars is seriously inconvenienced. This is the most obvious of the complications to simple rotation. In an illuminating survey of the history of the subject, Ekman (1993) notes that the precession was observed by Hipparchus, who, in about 125 BC, reported a measure of its rate in remarkable agreement with modern estimates.

To see how the precession arises, consider a small satellite in a circular orbit that is inclined to the equatorial plane. Since the Earth is not spherical, its gravitational force on the satellite is directed precisely towards the centre of the Earth only when the satellite is above the equator (or, of course, one of the poles if it is in a polar orbit). At intermediate latitudes the latitude variation of gravity imparts a torque tending to pull the satellite orbit towards the equatorial plane. The torque acts in a sense perpendicular to the angular momentum vector, causing a precession of the orbit. The points at which the orbit crosses the equatorial plane are referred to as nodes. As seen in space, not relative to the rotating Earth, they drift progressively in a prograde sense, eastwards for a satellite with a west-to-east component of orbital motion. The rate of this drift gives a precise measure of the Earth's ellipticity, as represented by the coefficient (Eq. 6.14)

$$J_2 = (C - A)/Ma^2 = 1.082\,626 \times 10^{-3}, \qquad (7.1)$$

where C, A are the moments of inertia about polar and equatorial axes and M, a are the mass and equatorial radius. There is a corresponding opposite torque on the Earth, but for a man-made satellite this is not significant.

The same principle applies to interactions with the Sun and Moon, but in these cases there is a noticeable effect on the Earth. The torque, proportional to $(C - A)$, acts on the Earth's rotational angular momentum, $C\omega$, so that the rate of the resulting precession provides a measure of the quantity known as the dynamical ellipticity:

$$H = (C - A)/C = 3.273\,79 \times 10^{-3} = 1/305.4567. \qquad (7.2)$$

Combining Eqs. (7.1) and (7.2) we obtain the moment of inertia coefficient

$$C/Ma^2 = J_2/H = 0.330\,698. \qquad (7.3)$$

This is a measure of the density profile of the Earth. For a uniform sphere the value would be 0.4 and the smaller value in Eq. (7.3) is a measure of the concentration of mass towards the centre. The moments of inertia of other bodies in the Solar System are mentioned in Eq. (1.17) and Table 1.2. The result in Eq. (7.3) was first used in modelling the Earth by K. E. Bullen in the 1930s and remains an important parameter that any model must match, in the same way as the total mass must agree with the observed mass (Section 17.6).

The dynamical ellipticity, H, appears also in the explanation of another phenomenon, the Chandler wobble (Section 7.3). This is a motion of the Earth that involves no interaction with any other body. It arises because the rotational axis departs by a small angle, α, normally about 0.15 arcsec (0.7 microradian) from the symmetry axis, which is the axis of maximum moment of inertia. This means that the rotational energy exceeds the energy of symmetrical rotation with the same angular momentum by an amount proportional to α^2. The resulting gyroscopic torque tends to turn the Earth so that its symmetry axis coincides with the axis of angular momentum (which remains fixed for the purpose of this discussion because we are considering an effect that is entirely internal to the Earth). The torque, acting on the angular momentum, causes a prograde precession of the rotational axis about the angular momentum axis that is apparent as a cyclic variation of latitude.

For a rigid Earth the wobble period would be $1/H$ days $= 305$ days (see Eq. (7.2)). This was the period sought for many years before the discovery, by S. C. Chandler in 1891, of the observed period, about 432 days. The difference is explained by the elastic deformation of the Earth, with accompanying responses of the oceans and core to the gyroscopic torque. The deformation partially adjusts the equatorial bulge towards symmetrical rotation, reducing the torque and so lengthening the period of the motion. Thus, the wobble provides a measure of the global average rigidity of the Earth, a check on the elasticity inferred from seismology (Chapter 17).

The wobble is damped with a time constant of about 30 years (Eq. (7.25)), which means that it must be continuously maintained. The mechanism for its maintenance has been a contentious issue for more than a century. Coupling to irregular core motions, atmospheric motions and earthquakes have all been repeatedly examined and found inadequate. We now have better evidence of a cause, involving interaction of the atmosphere and oceans, expressed primarily by ocean floor pressure variations (Gross, 2000). The 14-month Chandler wobble is superimposed on a slightly smaller 12-month variation, driven by seasonal mass re-distribution, and is apparent as a cyclic variation in the latitudes of observatories, with an amplitude that is a beat of the 14- and 12-month periods.

The astronomical observations that are the subject of this chapter have a long history. The relevance to various solid-Earth problems emerged more recently. As already mentioned, the determination of the dynamic ellipticity is particularly important, but other observations of interest include the elasticity of the Earth at the wobble frequency and evidence of coupling of the core to rotation of the mantle. Lambeck (1980) gave a comprehensive review of the several causes of rotational irregularity discussed in this chapter and the following one.

7.2 Precession of the equinoxes

The two principal terms in the external gravitational potential of the Earth are given by Eq. (6.13). The central, r^{-1}, term is dominant, but the second term is non-central, that is, it has a latitude dependence, due to the equatorial bulge. In addition to the central gravitational force $-m(\partial V/\partial r)$, exerted on a mass m at (r,ϕ), there is a torque $-m(\partial V/\partial \phi)$, with a corresponding equal and opposite torque exerted by the mass on the Earth. The magnitude of the torque is proportional to (m/r^3). For small bodies, such as man-made satellites, the only consequence is a regression of the nodes of the satellite orbits, that is, a precessional motion of the orbits about the equatorial plane (Section 9.2). But the torques exerted on the Moon and Sun are balanced by the torques exerted by these bodies on the Earth and cause the precession of the Earth's rotational axis. The process is almost non-dissipative, although the possibility that there is some precessional dissipation in the core is considered in Section 24.7, in connection with the power of the geomagnetic dynamo.

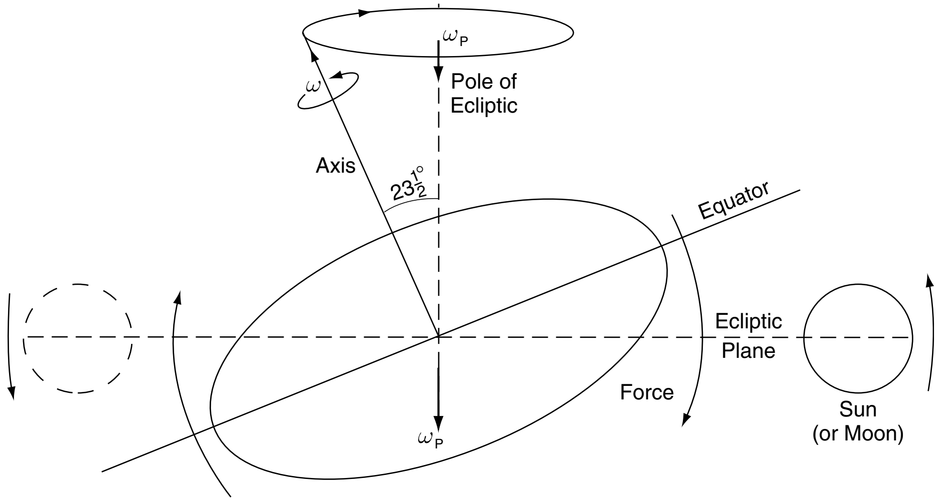


FIGURE 7.1 Origin of the precessional torque. The gravitational action of the Sun (and Moon) on the alignment of the Earth's equatorial bulge exerts a torque that tends to pull the bulge into alignment with the instantaneous Earth–Sun (or Earth–Moon) axis. The torque vanishes when the Sun (or Moon) crosses the equatorial plane, but appears with the same sign for both halves of the orbit, causing a net average precessional torque.

The cause of the precessional torque is illustrated in Fig. 7.1. The Moon and Sun pull the equatorial bulge towards alignment with themselves. The solar torque is a maximum at the solstices, when the Sun is 23.5° degrees from the equatorial plane, and it vanishes at the equinoxes, when the Sun is directly above the equator. But the sense of the torque is the same at both solstices, so that, although it occurs in semi-annual pulses, it has a cumulative effect. Similarly, the lunar torque occurs in semi-monthly pulses and its average effect is added to that of the Sun. The lunar contribution is slightly more than twice the solar one, because (m/r^3) is larger for the Moon.

Coupled with the precession are oscillatory motions of the axis towards and away from the pole of the ecliptic, that is, in a direction perpendicular to the precessional motion. They are referred to as nutations ('nodding'). Semi-annual and semi-monthly nutations arise from components of the solar and lunar torques perpendicular to the precessional component. These nutational torque components are zero at the solar and lunar solstices, when the torques are purely precessional, and at the equinoxes, when the torques vanish because the Sun or Moon are in the equatorial plane. A larger amplitude nutation (9.21 arcsec) of 18.6 year period arises from the coupling of the Earth to the precession of the pole of the lunar orbit about the pole of the ecliptic (Earth's orbital plane). The consequent variation in the inclination of the lunar orbit to the equator causes the Earth's nutation with the same period.

To examine quantitatively the precessional torque on the Earth, consider the geometry of Fig. 7.2, with the Sun, mass M_S, instantaneously at geocentric latitude ϕ and distance R from the Earth's centre. The gravitational potential at the centre of the Sun due to the Earth is given by Eq. (6.13), so that the torque exerted on the Sun, and therefore by the Sun on the Earth, is

$$L = M_S \frac{\partial V}{\partial \phi} = \frac{3GM_S}{R^3}(C - A)\sin\phi\cos\phi. \tag{7.4}$$

This torque acts about an axis in the equatorial plane normal to the Earth–Sun line, that is, it tends to pull the bulge into line with the instantaneous Earth–Sun axis. We can resolve it into components L_x about Ox and L_y about Oy. Then L_x is the torque component causing the semi-annual nutation and L_y is the solar precessional torque.

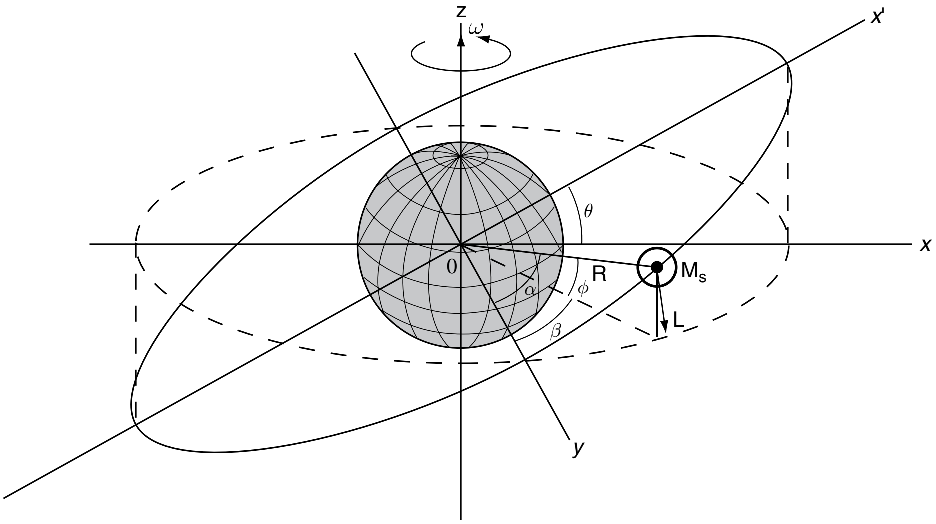


FIGURE 7.2 Geometry of the precessional torque. O*xy* represents the equatorial plane and *z* is the rotational axis. The Sun is at an instantaneous geocentric latitude ϕ, on an orbit of radius *R* in the plane O*x*′*y*, inclined at $\theta = 23.5°$ to the equatorial plane. The projection of the orbit onto the equatorial plane is shown as the dashed ellipse. For this purpose we can consider the orbit of the Sun about the Earth as equivalent to the orbit of the Earth about the Sun. A similar figure applies to the Moon.

$$L_y = \frac{3GM_S}{R^3}(C - A)\sin\phi\cos\phi\sin\beta, \tag{7.5}$$

where β is the longitude of the Sun, relative to its longitude at the equinox. Two trigonometric identities relate ϕ, β to θ, α, where α is the azimuthal angle of the Sun in its orbital plane and θ is here the maximum value of ϕ (23.5°):

$$\sin\phi = \sin\theta\sin\alpha, \tag{7.6}$$

$$\tan\phi = \tan\theta\sin\beta, \tag{7.7}$$

so that, in terms of θ and α,

$$L_y = \frac{3GM_S}{R^3}(C - A)\sin\theta\cos\theta\sin^2\alpha. \tag{7.8}$$

Since α gives the Sun's position in the orbital plane,

$$\overline{\sin^2\alpha} = 1/2, \tag{7.9}$$

and therefore the mean precessional torque, averaged over a year, is

$$\overline{L_y} = \frac{3}{2}\frac{GM_S}{R^3}(C - A)\sin\theta\cos\theta. \tag{7.10}$$

This annual average torque acts, in the sense of the arrows in Fig. 7.1, about the axis O*y* (Fig. 7.2).

Since O*y* is perpendicular to O*z*, which is the axis of the Earth's angular momentum, $C\omega$, the solar precessional torque causes an angular rate of change in the orientation of the rotational axis

$$\begin{aligned}\Omega_S = \overline{L_y}/C\omega &= \frac{3}{2}\frac{GM_S}{R^3}\frac{(C - A)}{C\omega}\sin\theta\cos\theta \\ &= \frac{3}{2}\frac{\omega_S^2}{\omega}\frac{C - A}{C}\sin\theta\cos\theta,\end{aligned} \tag{7.11}$$

where $\omega_S^2 = GM_S/R^3$ is Kepler's third law (Eq. (B.23), Appendix B), ω_S being the angular speed of the Earth–Sun orbital motion.

Still supposing the solar torque to act alone, as the Earth's rotational axis moves, the equatorial plane moves with it, causing the positions of the equinoxes (O*y* in Fig. 7.2) to move around the orbit. The axis of the precessional torque moves with O*y*, so that the torque causes the cyclic precession of the rotational axis about the pole of the ecliptic, describing a cone of semi-angle θ. If the solar torque acted alone, one cycle of the precession would take a time $\tau_{PS} = 2\pi/\omega_{PS} = 2\pi\sin\theta/\Omega_S$, where

$$\omega_{PS} = \Omega_S/\sin\theta = \frac{3}{2}\frac{\omega_S^2}{\omega}\frac{C - A}{C}\cos\theta. \tag{7.12}$$

A similar expression applies to the lunar contribution, ω_{PL} (except that in substituting Kepler's third law it is the mass of the Earth that is dominant for the Earth–Moon system), and the observed rate of precession, caused by both solar and lunar torques, is

$$\omega_{\mathrm{P}} = \omega_{\mathrm{PS}} + \omega_{\mathrm{PL}} = 50.3846(13)\ \text{arcsec/year}. \tag{7.13}$$

Since ω_{P} is well observed, as are the other quantities in Eq. (7.12) except the moments of inertia, the precession gives a measure of the dynamical ellipticity, $H = (C - A)/C$ (Eq. (7.2)). The fact that this is less than the geoid ellipticity demonstrates that the deeper, denser layers of the Earth, notably the core–mantle boundary, are less elliptical than the surface (see problem 6.4, Appendix J). Combining Eqs. (7.2) and (6.14) we obtain the coefficient of the Earth's moment of inertia, C/Ma^2 (Eq. (7.3)). This is an important parameter which Earth models are constrained to fit (Section 17.6).

Since the precession progressively changes the orientation of the rotation axis relative to the perigee and apogee of the orbit, there is an associated climatic cycle. There are also cyclic variations in the inclination of the ecliptic and the orbital eccentricity. These periodicities in orbital characteristics have climatic implications first studied in detail by M. Milankovitch in the early 1940s. Climatic cycles are observable in the sedimentary record at these periods (Berger, 1988), and Williams (1991) has found evidence of Milankovitch cycles as early as 440 million years ago. These observations provide a useful check on the evolution of the lunar orbit (Section 8.4). Implications for climate are considered in Section 26.4.

7.3 The Chandler wobble

Independently of its gravitational interactions with external bodies, the Earth undergoes a free, Eulerian precession, sometimes called the free nutation in geophysical literature, although it is not strictly a nutation. To distinguish it from the forced motions due to interactions with other bodies, the term wobble, or, in the name of its discoverer, Chandler wobble, is preferred. The wobble results from the rotation of the Earth about an axis that departs slightly from its axis of symmetry (or greatest moment of inertia). The total angular momentum remains constant, in magnitude and direction, but the symmetry axis follows a circular path about the spin axis, which remains nearly fixed in absolute orientation. The wobble is apparent as a cyclic variation of latitude with a period of about 432 days (1.2 years) and variable amplitude averaging about 0.15 arcsec (Vondrák, 1999). The Chandler wobble is superimposed on a 12-month (seasonal) latitude variation of 0.1 arcsec amplitude, so that the two periods are observed to beat, resulting in a six-year cycle in the amplitude of the latitude variation (Fig. 7.3).

The theory of the wobble period is usually presented in terms of Euler's equations, but greatest geophysical interest arises from the mechanisms of excitation and damping and, for this purpose, it is convenient to use a derivation that makes use of wobble energy.

Consider a hypothetical, rigid Earth with principal moments of inertia C, A, A, where $C > A$, rotating with angular velocity ω about an axis at a very small angle, α, to the C axis. The total rotational energy is the sum of the energies of the components of rotation about the three principal axes,

$$E_{\mathrm{T}} = \frac{1}{2}\left(Cm_3^2 + Am_1^2 + Am_2^2\right)\omega^2, \tag{7.14}$$

where m_1, m_2, and m_3 are the direction cosines of the rotational axis to the A, A and C axes. Thus

$$m_1^2 + m_2^2 = \alpha^2, \tag{7.15}$$

where $m_1^2 + m_2^2 + m_3^2 = 1$.

The energy of rotation with the same angular momentum about the C axis, that is, with $\alpha = 0$ and no wobble, is

$$E_0 = \frac{1}{2}C\omega_0^2. \tag{7.16}$$

Thus, the kinetic energy of the wobble is the difference between Eqs. (7.14) and (7.16). Making use of Eq. (7.15), this is

$$E_{\mathrm{W}} = E_{\mathrm{T}} - E_0 = \frac{1}{2}C\omega^2\left[1 - \left(\frac{C - A}{C}\right)\alpha^2\right] - \frac{1}{2}C\omega_0^2. \tag{7.17}$$

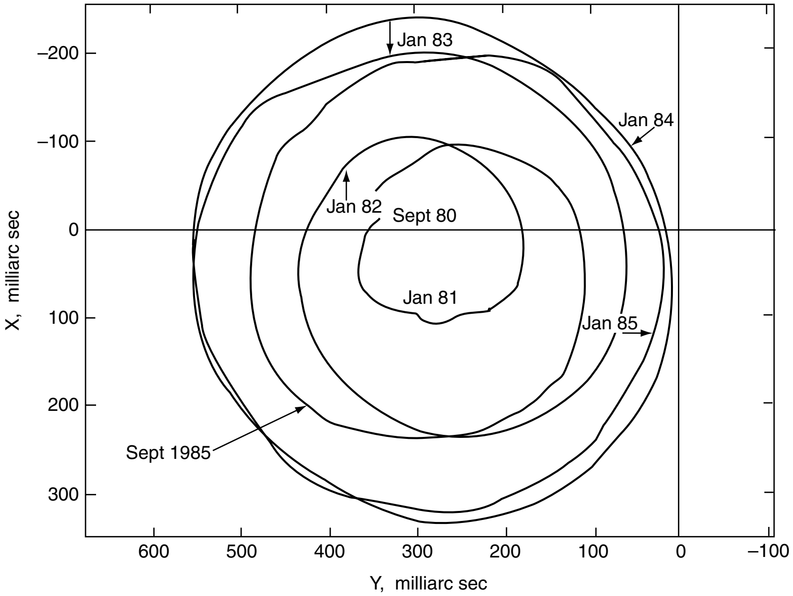


FIGURE 7.3 Path of the pole from late 1980 to late 1985, approximately one beat period of the combined 432- (Chandler) and 365-day (seasonal) motions. *X* and *Y* represent displacements in the direction of the Greenwich meridian and 90° East. Accuracy of the observations improved during this period. By 1985 it was clear that the irregularities in the pole path are real. From a record presented by Carter (1989).

The angular momentum is the vector sum of its components about the principal axes and is conserved, so that it is equal to $C\omega_0$ (with no wobble), that is

$$C\omega_0 = \left(C^2m_3^2 + A^2m_1^2 + A^2m_2^2\right)^{1/2}\omega$$
$$= \left[1 - \left(\frac{C^2 - A^2}{C^2}\right)\alpha^2\right]C\omega. \quad (7.18)$$

Substituting for ω_0 by Eq. (7.18) in Eq. (7.17), we have

$$E_W = \frac{1}{2}A\left(\frac{C-A}{C}\right)\omega^2\alpha^2 = \frac{1}{2}AH\omega^2\alpha^2, \quad (7.19)$$

where H is the dynamical ellipticity (Eq. (7.2)) obtained from the precession.

Since the excess rotational energy is a function of the angle α, it follows that there is a gyroscopic torque, L_g, that tends to turn the Earth to its lowest energy state ($\alpha = 0$):

$$L_g = -\frac{dE_W}{d\alpha} = -AH\omega^2\alpha. \quad (7.20)$$

This torque acts on the angular momentum component in the equatorial plane, $A\alpha\omega$, that is, the component associated with the axial misalignment, and, for the hypothetical rigid Earth, causes free precessional motion at an angular frequency

$$\omega_W = -\frac{L_g}{A\omega\alpha} = H\omega. \quad (7.21)$$

This is a prograde motion with a period $C/(C-A)$ days $= 305$ days.

A 305-day latitude variation was sought for many years before the discovery of the 432 day period by S. C. Chandler in 1891. The difference is primarily due to the elastic deformation of the Earth, which partially accommodates the equatorial bulge to the instantaneous rotation axis, reducing the gyroscopic torque. There are also smaller effects of the oceans and core. Neglecting the lesser contributions, the solid deformation reduces the gyroscopic torque, L_g, by the factor $305/432 = 0.7$, which means that the bulge is deflected by an angle 0.3α, requiring a shear strain

$$\varepsilon = 0.3\alpha f \approx 1 \times 10^{-3}\alpha, \quad (7.22)$$

where $f \approx 1/300$ is the equatorial flattening.

As mentioned in Section 7.1, for more than a century a satisfactory explanation for the excitation of the wobble was elusive and the same is true of its damping. If the excitation is random then the observed spectrum of the wobble resonance gives a measure of the damping and this is expressed in terms of the Q of the wobble, for which there are two alternative (but formally equivalent) definitions:

$$Q_W = -2\pi E_W/\Delta E_W, \tag{7.23}$$

$$Q_W = \omega_W/\delta\omega_W. \tag{7.24}$$

Here $-\Delta E_W$ is the diminution in one wobble cycle of the energy, E_W, of an unmaintained wobble, not a directly observed quantity. $\delta\omega_W$ is the spectral line width (at half power) of the wobble, so that Q_W is defined as the sharpness of the wobble frequency. These definitions have equivalents in electrical circuit theory, where the term Q – quality factor – originates. Jeffreys (1959) presented a mechanical analogue that is easy to visualize – a circular pendulum and a group of boys with pea-shooters around it, randomly firing peas at it. The pendulum swings in response to the impulses and if it is only lightly damped it may develop a large amplitude, so that individual impulses have only slight effects. This is the high Q situation with little dissipation of energy per cycle and a sharply defined period of oscillation. If the Q is lowered by additional damping, the amplitude is reduced, individual impulses become proportionately more important and the period becomes more blurred. Application of Eq. (7.24) to observations gives

$$Q_W \approx 80, \tag{7.25}$$

corresponding to decay of an unmaintained wobble with a time constant $2Q_W/\omega_W \approx 30$ years, but with a wide range of uncertainty ($30 < Q_W < 180$). An analysis by Vondrák (1999) suggests that the wobble is not a simple linear phenomenon, which the concept of Q assumes, but that the amplitude and period are related. However, this may be an artifact of the excitation mechanism.

The role of mantle anelasticity in damping the wobble can be examined by considering the anelastic Q required to give the observed wobble Q (Eq. (7.25)) with a strain amplitude given by Eq. (7.22). The volume-averaged rigidity modulus of the Earth is $\mu \approx 124$ GPa, so that the elastic strain energy for a wobble of amplitude α is

$$E_S \approx (1/2)\mu\varepsilon^2 V = 6.7 \times 10^{25}\alpha^2 \text{ Joules}, \tag{7.26}$$

where V is the volume of the Earth. This compares with

$$E_W = 7.0 \times 10^{26}\alpha^2 \text{ Joules} \tag{7.27}$$

by Eq. (7.19). Thus, for mantle anelasticity to damp the wobble would require an anelastic $Q_S \approx Q_W/10 \approx 8$ and this is so much smaller than the seismologically observed Q_S as to be implausible. The mantle can make only a minor contribution to Q_W.

It is important to note that an excitation event becomes ineffective if it extends over more than half a wobble period (7 months). Internal torques or mass displacements that move the symmetry axis away from the rotation pole at one instant would move it closer half a period later, when the symmetry axis is on the opposite side of the pole. This leaves coupling of the Earth to the atmosphere and oceans as the only viable mechanism, as concluded by Gross (2000). A suggestion that the atmosphere alone suffices is kept alive by Aoyama and Naito (2001), although their argument is based on evidence for a sharp 14-month peak in the spectrum of atmospheric (wind plus pressure) excitation, which is hard to understand. A reasonably sharp 'line' spectrum of excitation for the annual wobble can be understood as a seasonal effect, so that in this case the line width is unrelated to damping, unlike the case of random excitation.

The existence of the forced annual wobble demonstrates that seasonal movements of the atmosphere and hydrosphere cause an annual polar shift of ±0.05 arcsec. The 14-month wobble, of average amplitude about 0.15 arcsec, with dissipation at a rate corresponding to a free decay time constant of 25 periods, requires excitation by a random source of amplitude $0.15/\sqrt{25} = 0.03$ arcsec, a surprisingly large fraction of the annual excitation. But the wobble Q is not well observed and these numbers are insecure.

Superimposed on the wobble there is a steady drift of the pole towards 79° W, at a rate which Vondrák estimated to be 0.351″/century (11 cm/year). It is attributed to asymmetry of the mass redistribution caused by post-glacial rebound. The total displacement of the pole in this process is only of order 1 km. It is not an observation of polar wander, but it is possible in principle that true polar wander, driven by mass displacements in the mantle accompanying changes in convection pattern, is also occurring.

7.4 Length-of-day (LOD) variations

On a geological time scale, the major variation in the rate of rotation of the Earth is a gradual slowing by tidal friction (Section 8.3), which converts rotational angular momentum to angular momentum of the Earth–Moon and Earth–Sun orbits. This is a permanent and irreversible process, but is sufficiently slow to be obscured by shorter term angular momentum exchanges with the atmosphere, oceans and core. There is, at the present time, a component of angular acceleration resulting from post-glacial rebound, which causes small changes in the axial moment of inertia. The more rapid fluctuations are now clearly identified with the atmosphere. Figure 7.4 shows variations of 2 ms in the length of day, that is 2.3 parts in 10^8, closely correlated with changes in atmospheric angular momentum. Since the moment of inertia of the atmosphere is only 1.7×10^{-6} of the total for the Earth (Table A.4, Appendix A), the extreme fluctuations in the average rate of rotation of the atmosphere exceed 1%. This corresponds to a change in global average wind of order 5 m s^{-1} or 18 km/hour. But, as the broken curve of Fig. 7.4 indicates, such rapid changes in the atmosphere occur regularly, with a strong annual cycle and shorter period variations as well as apparently random fluctuations. The correlation with LOD observations (the continuous curve of Fig. 7.4) demonstrates that rapid LOD variations must be attributed to a tight coupling of the solid Earth to atmospheric fluctuations. The energy implications of this coupling are considered in Section 26.2.

Somewhat larger LOD fluctuations, that occur on a time scale of decades, are interpreted in terms of angular momentum exchange with the core. Holme and deViron (2005) reported evidence that LOD variations are correlated with rapid changes in the geomagnetic secular variation (jerks – see Section 24.3), implying tight core–mantle coupling as well as rapid changes in core angular momentum. The moment of inertia of

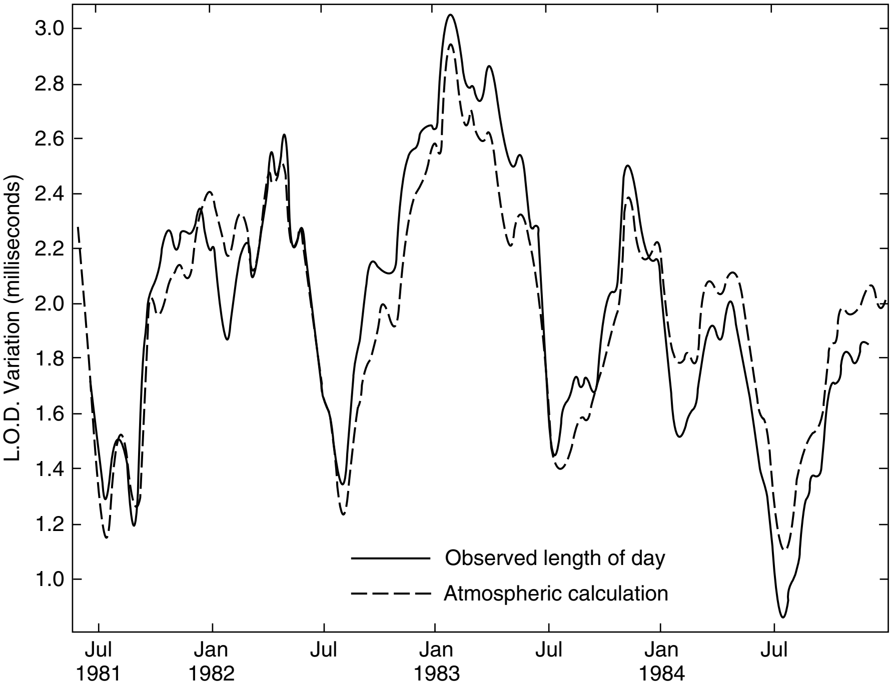


FIGURE 7.4 Changes in the length of day observed by very long baseline interferometry (VLBI) compared with the changes expected from variations in the angular momentum of the atmosphere. Redrawn from Carter (1989).

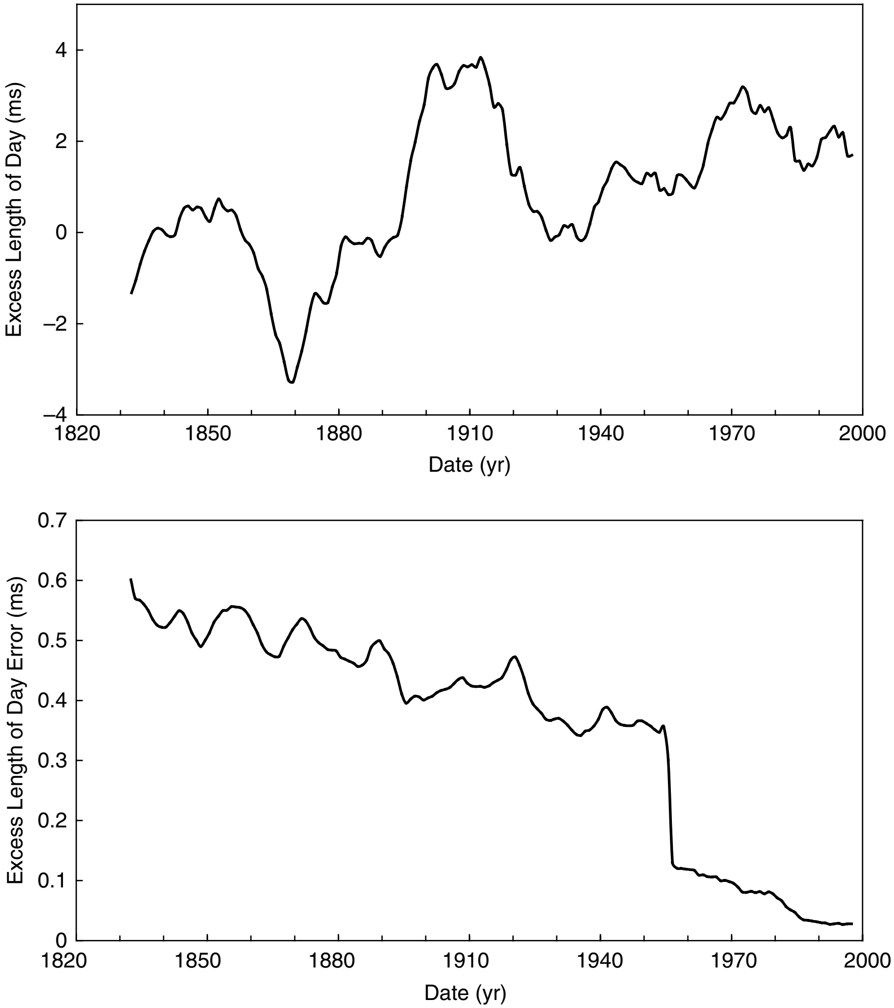


FIGURE 7.5 The decade variations in the length of the day from an analysis by Gross (2001), data supplied by R. S. Gross. The lower figure shows the uncertainty, which was sharply reduced by the introduction of quartz crystal clocks in the 1950s.

the core is 13% of that of the mantle so a change of 4 ms in the LOD (4.6 parts in 10^8), which is seen to be typical of the record in Fig. 7.5, implies a change of 3.6 parts in 10^7 in the average rate of rotation of the core, corresponding to a speed of 10^{-4} m s^{-1} (3 km/year) at the core surface. This is almost 10% of the speed of core motions inferred from the geomagnetic secular variation (Section 24.3) and indicates rapid coordinated changes in large volumes of core fluid. The mechanism of core–mantle coupling is of interest partly because of its relevance to other phenomena, especially the geomagnetic westward drift (Section 24.6). The observations themselves do not allow us to say whether the decade time scale is characteristic of this coupling or of the core motions, allowing a much tighter coupling, but either way it is clear that major changes in core motions occur in a decade or so. The spectrum of length of day fluctuations plotted by Gross (2001) indicates a relaxation time of about 8 years, and in the following section this is assumed to be characteristic of the coupling.

The slow decrease in the LOD caused by tidal friction, a subject of Chapter 8, is effective over a geological time scale. At the other extreme, human activity, especially the impoundment of water in reservoirs, occurs on a decade time scale.

It has lowered sea level by several centimetres (see discussions by Chao *et al.*, 1994), with effects on the moment of inertia, and hence the length of day, by amounts that are too small to observe.

7.5 Coupling of the core to rotational variations

Consider a simple situation in which the core rotates coherently at angular speed $(\omega + \Delta\omega_c)$ and the mantle rotates at $(\omega + \Delta\omega_m)$, where ω is their common rate at equilibrium. Then, to conserve angular momentum

$$I_m \Delta\omega_m + I_c \Delta\omega_c = 0, \tag{7.28}$$

where $I_m = 8.04 \times 10^{37}\,\mathrm{kg\,m^{-3}}$, and $I_c = 0.92 \times 10^{37}\,\mathrm{kg\,m^{-3}}$, are the moments of inertia of the mantle and core. The relative angular velocity is

$$\begin{aligned}\Delta\omega = \Delta\omega_m - \Delta\omega_c &= \Delta\omega_m(1 + I_m/I_c)\\ &= 8.74\Delta\omega_m.\end{aligned} \tag{7.29}$$

If the coupling is linear, so that the mutual torque, L, restoring equilibrium, is proportional to $\Delta\omega$, then we can define a coupling coefficient, K_R, such that

$$L = I_m \frac{\mathrm{d}}{\mathrm{d}t}(\Delta\omega_m) = -K_R\Delta\omega = -K_R\left(1 + \frac{I_m}{I_c}\right)\Delta\omega_m. \tag{7.30}$$

This integrates to

$$\Delta\omega_m = (\Delta\omega_m)_o \exp(-t/\tau), \tag{7.31}$$

where the relaxation time is

$$\tau = \left[K_R\left(\frac{1}{I_m} + \frac{1}{I_c}\right)\right]^{-1}. \tag{7.32}$$

If we take $\tau = 8$ years, then

$$K_R \approx 3.2 \times 10^{28}\,\mathrm{N\,m\,s\,(kg\,m^2\,s^{-1})}, \tag{7.33}$$

which is a plausible value for electromagnetic coupling. There may also be topographic coupling caused by the flow of core fluid across undulations of the core–mantle boundary, but we have no observations to indicate how strong it may be. Boundary topography is obscured to seismological observation by heterogeneities at the base of the mantle. This estimate of K_R must be regarded as a lower bound, because it assumes the 8-year time constant to be a characteristic of the coupling mechanism, not the core motions. Doubt about the validity of Eq. (7.33) arises from its neglect of a gravitational interaction between the mantle and the inner core. Length of day variations by a few parts in 10^8 are small compared with the differential rotation of the core and mantle suggested by the geomagnetic westward drift, about 1 part in 10^6. These effects may all be controlled by gravitational coupling to the inner core, as discussed in Section 24.6.

Assuming that the moments of inertia of the mantle and core are unaffected by the rotational changes, the excess energy, relative to coherent rotation, is

$$\begin{aligned}\Delta E &= (1/2)I_m[(\omega + \Delta\omega_m)^2 - \omega^2]\\ &\quad + (1/2)I_c[(\omega + \Delta\omega_c)^2 - \omega^2]\\ &= (1/2)I_m(\Delta\omega_m)^2 + (1/2)I_c(\Delta\omega_c)^2\\ &= (1/2)I_m(1 + I_m/I_c)(\Delta\omega_m)^2.\end{aligned} \tag{7.34}$$

For a 4 ms departure of the LOD from equilibrium, $\Delta\omega_m = 4.63 \times 10^{-8}\omega = 3.38 \times 10^{-12}\,\mathrm{rad\,s^{-1}}$, giving $\Delta E = 4.46 \times 10^{15}$ Joule, implying a dissipation of 1.8×10^7 W for an 8-year relaxation time. This is not significant to core–mantle boundary processes.

Having estimated the coupling coefficient, K_R, from LOD observations (Eq. (7.33)), we can consider the consequences of assuming that it applies also to the Chandler wobble and to the precession. Consider first the wobble. To maximize the supposed effectiveness of the core in either damping or exciting the wobble, imagine the mantle to be wobbling (at the Chandler frequency, ω_W) with angular amplitude α over a non-wobbling core. Then the relative angular velocity of the mantle and core is

$$\Delta\omega_W = \omega_W\alpha. \tag{7.35}$$

The resulting mutual torque is

$$L_W = -K_R\Delta\omega_W \tag{7.36}$$

and the consequent energy dissipation is

$$-\frac{\mathrm{d}E_W}{\mathrm{d}t} = -L_W\Delta\omega_W = K_R\omega_W^2\alpha^2. \tag{7.37}$$

Comparing this with the wobble energy for the simple, rigid Earth model given by Eq. (7.19), the

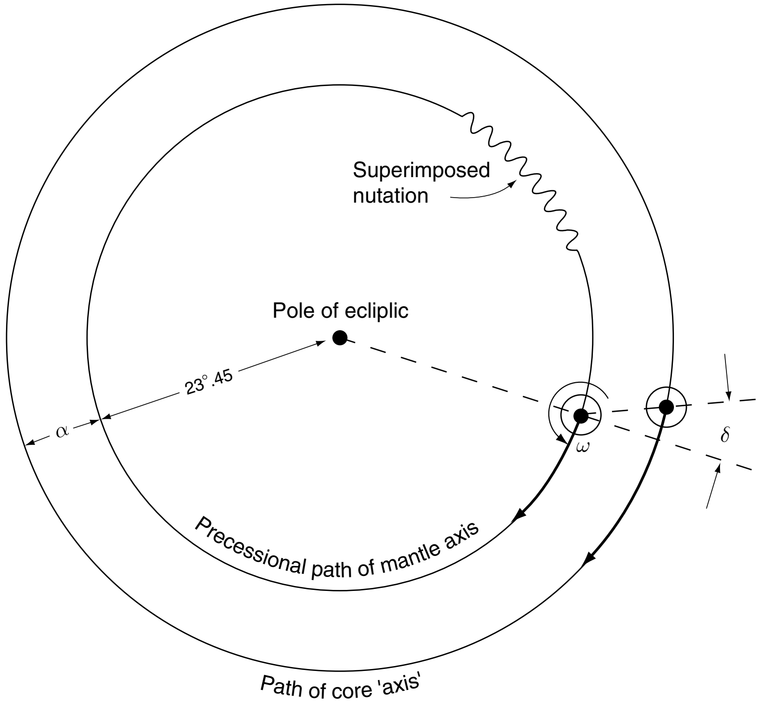


FIGURE 7.6 Precessional paths of the mantle and core axes resulting from the simple model of inertial coupling of the core and mantle. The precessional motion is retrograde (opposite to the Earth's axial rotation, ω) and the superimposed nutation is shown only for part of the path.

relaxation time for decay of the wobble amplitude by this coupling is

$$\begin{aligned}\tau_W &= \frac{2E_W}{(-\mathrm{d}E_W/\mathrm{d}t)}\\ &= \frac{AH}{K_R}\left(\frac{\omega}{\omega_W}\right)^2 = 1.5\times 10^{12}\ \mathrm{s}\\ &= 48\,000\ \text{years}. \end{aligned} \tag{7.38}$$

This is twice the time constant for decay of energy, which varies as the square of amplitude. It is so much longer than the observed relaxation time of about 30 years that it can have no relevance to wobble damping.

Now consider the precession. As pointed out in Section 6.4, the decrease in the ellipticity of the Earth's mass with depth is a consequence of the increase in density with depth. Of particular interest in this connection is the ellipticity of the core. Having a smaller dynamical ellipticity than the Earth as a whole, it is not subject to a lunisolar precessional torque of sufficient magnitude to cause it to keep up with the precession. (But we note that Mathews *et al.* (2002) reported evidence from nutations that the core is more elliptical than equilibrium theory suggests.) It must be very effectively coupled to the mantle, because semi-independent precession of the core and mantle would lead to widely different rotational axes and velocity differences of hundreds of metres per second across a boundary layer at the core–mantle interface, some million times faster than the speeds indicated by geomagnetic secular variation (Section 24.3). The theory of the dominant coupling mechanism, termed inertial coupling, originated with H. Poincaré. The principle can be explained in terms of the following simple model suggested by A. Toomre.

Consider a frictionless particle sliding around an oblate ellipsoidal cavity free of gravity, initially following the equatorial plane. When the axis of the cavity is turned through a small angle, the particle continues to orbit in the original plane, but as this is now inclined to the equator of the cavity the orbit has become slightly elliptical. Moreover, the cavity wall is no longer everywhere perpendicular to the plane of the orbit and, since the particle is frictionless, it experiences only a force normal to the cavity wall. This is not a central force; there is a component perpendicular to the plane of the orbit. The result is a torque that causes the orbit to precess in a retrograde sense about the equator of the cavity. This is a simplified analogue of the response of the fluid core motion to a change in orientation of the axis of the core–mantle boundary. The important simplification is neglect of the internal motion in the core (Poincaré flow) that is required to accommodate its shape to an ellipsoidal cavity misaligned with its own rotational axis.

Figure 7.6 illustrates how the torque exerted by the mantle on the core keeps the core in step with the precession. The tendency of the core to be left behind gives the precessional coupling to the mantle, so that the core axis 'tries' to precess about the mantle axis. But with the continued precession of the mantle, this serves simply to keep the core in step, with a small angular separation, α, of the axes. A value of α of about 6×10^{-6} rad (1.2 arcsec) gives the required torque on the core (Stacey, 1973).

Inertial coupling is non-dissipative, and if it accounted completely for the locking of the core to the precession, then the phase lag, δ in Fig. 7.6, would be zero. Although the relative motion of the core and mantle is reduced by the Poincaré flow, it is not zero, and so electromagnetic coupling between them ensures that some dissipation occurs. The Poincaré flow deforms the magnetic field lines in the core with consequent ohmic dissipation. The magnitude of this contribution to dynamo power is

$$\mathrm{d}E/\mathrm{d}t = -K_\mathrm{R}(\alpha\omega)^2 = 6.3\times 10^9\ \mathrm{W}, \qquad (7.39)$$

where $K_\mathrm{R} = 3.2\times 10^{28}\,\mathrm{kg\,m^2\,s^{-2}}$ is the core–mantle coupling coefficient adopted from the length of day calculations (Eq. 7.33) and $\omega = 7.29\times 10^{-5}\,\mathrm{rad\,s^{-1}}$ is the rotation rate. This is not a significant contribution to present dynamo power, but precession was stronger in the distant past, when the Moon was closer and faster rotation caused greater ellipticity. So, we examine its relevance to the dynamo early in the life of the Earth. Neglecting any change in the obliquity (θ), Eq. (7.11) gives a precession rate proportional to $(C-A)/R^3\omega$, with $(C-A)\propto\omega^2$ (see Eq. 6.19), so that the rate of precessional motion of the rotational axis is proportional to ω/R^3.

The core rotational axis 'tries' to precess about the cavity axis at an angular rate proportional to $\omega\varepsilon$, where $\varepsilon\propto\omega^2$ is the cavity ellipticity, giving a relative precession rate proportional to ω^3. The angle of axial misalignment, α, is self-adjusted so that this differential precession keeps the core in step with the precession of the mantle, that is $\alpha\propto(\omega R^3)/\omega^3 = 1/R^3\omega^2$. Equation. (7.39) gives dissipation proportional to $(\alpha\omega)^2$, that is to$1/R^6\omega^2$. Evolution of the lunar orbit, discussed in Section 8.4, indicates that the Moon was probably initially at about $R = 30$ Earth radii (half of the present distance). In that case the rotation rate, ω, would then have been 2.5 times the present value and with these factors the dissipation would have been 10 times the present rate, that is 6×10^{10} W. It is included in the core energy budget, in Section 21.4 and Table 21.5, and its significance to the dynamo early in the life of the Earth is discussed in Section 24.7.

8

Tides and the evolution of the lunar orbit

8.1 Preamble

The Earth rotates in the gravity fields of the Moon and Sun, causing cyclic variations in the gravitational potential. The most obvious consequence is the marine tide but there are also tidal deformations of the solid Earth. These take the form of prolate ellipsoidal extensions, aligned with the Moon or Sun, and the observed tides are caused by the rotation of the Earth in the deformed 'envelope'. This phenomenon occurs in all rotating astronomical bodies that are not very remote from the gravitational influences of their neighbours. It would merit little more than a footnote in a text of this kind were it not for one crucial effect: tides are dissipative. The dissipation of rotational energy is slight in gaseous bodies, but it strongly influences the behaviour and orbital evolution of solid planets and satellites and in the Earth it is strongest in the oceans.

We present an analysis of tidal potential and deformation as the starting point for a discussion of tidal friction. This has had a major effect on the Earth–Moon system and we suggest also on Venus and Mercury. It may not be important to Mars but it has interesting effects on some of the solid bodies of the outer Solar System. The vigorous volcanism of Jupiter's satellite Io is attributed to heating by tidal dissipation. In this case it is a radial tide, that is, a variation in the amplitude of a tide of constant orientation, because Io has stopped rotating relative to Jupiter. The ellipticity of its orbit is maintained by interactions with other satellites. Pluto and its major satellite, Charon, appear to present fixed faces to one another, presumably because tidal friction has stopped their relative rotations, just as the Moon presents a fixed face to the Earth.

The outer planets have many satellites but the four terrestrial planets have just three between them, two small ones orbiting Mars and Earth's large one. There are none at all orbiting either Venus or Mercury. Tidal friction allows a straightforward explanation for this situation. The inner planets and the Moon are solid and their tides dissipate rotational energy. This has the effect of slowing and eventually stopping relative rotations. The large tide raised in the Moon by the Earth has completely stopped its rotation relative to the Earth. The tide raised by the Moon in the Earth is slowing the Earth's rotation with a consequent transfer of angular momentum to the lunar orbit, causing the Moon's orbit to expand at a rate that is currently about 3.8 cm/year. Over geological time this means a dramatic change in the orbit, as discussed in Section 8.4. In Section 8.6 we present the case for two moons surviving independently for the first 600 million years and being brought together by tidal friction 3.9 billion years ago.

Venus and Mercury present a different situation. Being much closer to the Sun they have strong solar tides, which have so slowed their rotations that they present almost constant faces to the Sun. Any satellites would have been orbiting faster than the planetary rotations, with tidal friction opposing their motions and causing them to spiral inwards. Orbital angular momentum, transferred to planetary rotation, would be lost to the solar tidal friction. Thus, for these planets any early satellites have merged with the parent planets. It is important to note the very strong dependence of tidal friction on the separation, r, of interacting bodies. Tidal amplitude varies as r^{-3} and dissipation as the square of amplitude, that is r^{-6}, or even more strongly if dissipation is non-linear. The motion of a satellite spiralling in towards its parent planet would accelerate rapidly if tidal friction is effective at all.

Tides are ubiquitous in planetary and satellite systems and tidal dissipation in solid bodies, such as the terrestrial planets, has not received attention commensurate with its importance to the evolution of the inner Solar System. This chapter draws attention to the need to give it greater recognition in Earth and planetary science.

8.2 Tidal deformation of the Earth

Consider the geometry of Fig. 8.1, in which the Earth and Moon are orbiting about their common centre of mass, which is at distance b from the centre of the Earth. For the moment, consider the Earth to present a constant face to the Moon, that is it has an axial rotation as well as orbital motion at angular speed ω_L. Thus the whole figure rotates, with P at fixed distance from the axis. We wish to calculate the potential at P due to the gravity of the Moon plus the consequent orbital motion. This is

$$W = -\frac{Gm}{R'} - \frac{1}{2}\omega_L^2 r^2. \tag{8.1}$$

Using the cosine rule

$$(R')^2 = R^2 + a^2 - 2aR\cos\psi \tag{8.2}$$

we have, to second order in the small quantity (a/R),

$$(R')^{-1} = R^{-1}\left(1 - \frac{1}{2}\frac{a^2}{R^2} + \frac{a}{R}\cos\psi + \frac{3}{2}\frac{a^2}{R^2}\cos^2\psi + \cdots\right). \tag{8.3}$$

We also have the trigonometric relationships

$$\cos\psi = \sin\theta\cos\lambda, \tag{8.4}$$

$$\begin{aligned} r^2 &= b^2 + (a\sin\theta)^2 - 2b(a\sin\theta)\cos\lambda \\ &= b^2 + a^2\sin^2\theta - 2ba\cos\psi \end{aligned} \tag{8.5}$$

and

$$b = \frac{m}{M+m}R, \tag{8.6}$$

as well as Kepler's third law (for which we generalize Eq. (B.23) in Appendix B to non-negligible m)

$$\omega_L^2 R^3 = G(M+m). \tag{8.7}$$

Substituting in Eq. (8.1) for R', r and b by Eqs. (8.3), (8.5) and (8.6) and then using Eq. (8.7) to simplify the result, we obtain an expression for W with three readily identified terms:

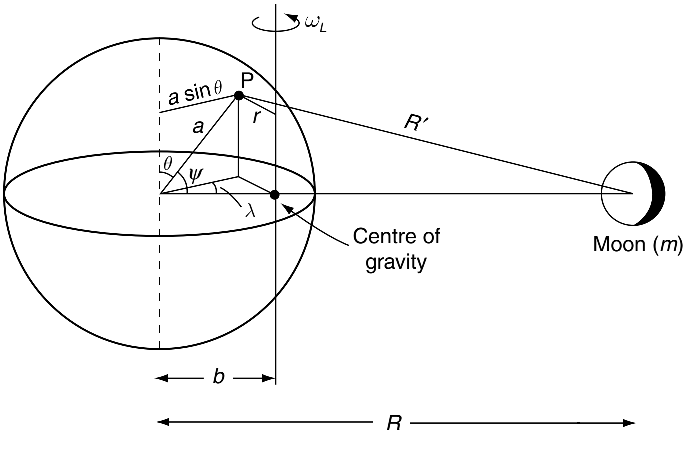


FIGURE 8.1 Geometry for calculation of tidal potential of the Moon (mass m) at distance R from the centre of the Earth and R' from the arbitrary point P on the surface of the Earth. The intersection of the lunar orbital plane with the Earth is shown as an ellipse and the angular coordinates of P are θ, referred to the normal to this plane, and λ, measured within the plane from the Earth–Moon axis. The centre of gravity of the system, about which both the Earth and Moon orbit, is at a distance b from the centre of the Earth, slightly less than the Earth radius, a.

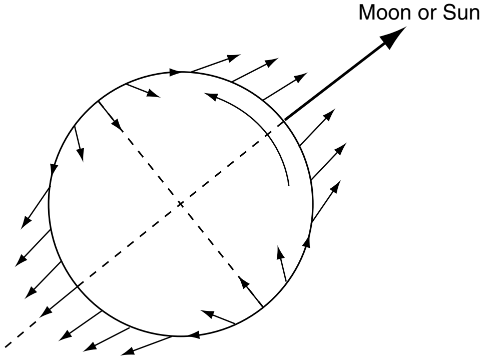


FIGURE 8.2 Tidal force at the surface of the Earth. This is the gradient of the tidal potential, W_2 (Eq. (8.9)).

$$\begin{aligned} W = &-\frac{Gm}{R}\left(1+\frac{1}{2}\frac{m}{M+m}\right) \\ &-\frac{Gma^2}{R^3}\left(\frac{3}{2}\cos^2\psi-\frac{1}{2}\right)-\frac{1}{2}\omega_{\mathrm{L}}^2a^2\sin^2\theta. \end{aligned} \tag{8.8}$$

The first term is the gravitational potential at the centre of the Earth due to the Moon, with a small correction. This term is a constant, independent of the position of P on the Earth and so has no tidal effect. The third term is the rotational potential at P due to rotation of the Earth about its own centre at speed ω_{L}. It is therefore simply contributory to the potential due to rotation of the Earth at angular speed $\omega \gg \omega_{\mathrm{L}}$, as in Eq. (6.15). The second term in Eq. (8.8) is the tidal potential. It is a second-order zonal harmonic and represents the deformation of an equipotential surface to a prolate ellipsoidal form, aligned with the Earth–Moon axis (Fig. 8.2). It is conventional to represent the tidal potential by its own symbol,

$$W_2 = -\frac{Gma^2}{R^3}\left(\frac{3}{2}\cos^2\psi-\frac{1}{2}\right), \tag{8.9}$$

From this equation we can see the origin of the tidal contribution to the oblateness of the Earth, mentioned in Section 6.1. At the pole of the lunar orbit, $\psi = 90^\circ$ and W_2 has a fixed value $Gma^2/2R^3$. In the orbital plane (approximately the equator), W_2 oscillates over the range $(Gma^2/2R^3)\left(-\frac{1}{2}\pm\frac{3}{2}\right)$, with an average value of $-Gma^2/4R^3$, the average value of $\cos^2\psi$ being ½. The difference between the polar and average equatorial values gives a contribution to the 'excess' ellipticity of the Earth, but this contribution cannot be regarded as a departure from hydrostatic equilibrium.

W_2 is a disturbance to the static potential of the Earth (mass M, radius a),

$$W_0 = -\frac{GM}{a}, \tag{8.10}$$

so that for the lunar tide

$$\left(\frac{W_2}{W_0}\right)_{\max} = \frac{m}{M}\left(\frac{a}{R}\right)^3 = 5.6\times 10^{-8}, \tag{8.11}$$

The corresponding quantity for the solar tide is 0.45 times this value. Still considering a hypothetical, rigid Earth, the tidal variation in gravity is

$$\delta g = \frac{\partial W_2}{\partial a} = \frac{Gma}{R^3}\left(3\cos^2\psi - 1\right), \tag{8.12}$$

the fractional change being

$$\frac{\delta g}{g} = -\frac{m}{M}\left(\frac{a}{R}\right)^3\left(3\cos^2\psi-1\right). \tag{8.13}$$

This is within the range of gravity meters used for geophysical surveying and tidal corrections are routinely applied to observations. The variation in height of the equipotential surface is, with no allowance for the effect of tidal deformation of the Earth,

$$\delta a = \frac{W_2}{g} = \frac{m}{M}\left(\frac{a}{R}\right)^3 a\left(\frac{3}{2}\cos^2\psi-\frac{1}{2}\right), \tag{8.14}$$

which has a peak-to-trough amplitude of 0.535 m for the lunar tide.

The lunar and solar tidal bulges are superimposed, with a relative alignment that changes progressively through the lunar month. Their effects are added at new and full Moon, but oppose when the Moon and Sun are 90° apart. The solar tide is smaller by the factor 0.46, so that the total semi-diurnal tidal amplitude varies roughly by the factor $(1+0.46)/(1-0.46) = 2.7$ between spring (maximum) and neap tides. The variation is apparent as a beat of the 12.42 hour and 12.00 hour periods. There are many other periodicities, due to the eccentricities of the orbits, the misalignment of the lunar and solar orbital planes, the precession of the lunar orbit and, most importantly, the inclination of the ecliptic (the misalignment of equator and the

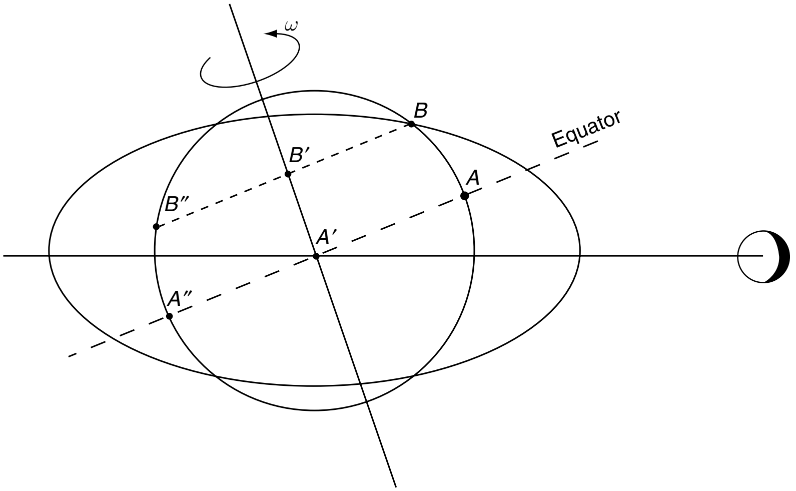


FIGURE 8.3 Rotation of the Earth about an axis that is inclined to the lunar and solar planes introduces an asymmetry to the tides – the tidal inequality – which appears as diurnal tidal components. A point on the equator, with successive positions *A*, *A*′, *A*″, sees a semi-diurnal tide, but the point following the path *B*, *B*′, *B*″ sees a prominent diurnal tide.

orbital plane). The tidal inequality arising from this misalignment gives lunar and solar diurnal components to the tides (Fig. 8.3).

The deformation of both the solid Earth and the oceans modifies the tidal potential. The deformation assumes the form of W_2 and so can be described by dimensionless factors, h and k, which were introduced by A. E. H. Love and are known as Love numbers, and a third, l, suggested by T. Shida, defined as follows, with numerical values for ellipsoidal deformation of the Earth (Mathews *et al.*, 1995).

$h = 0.603$ is the ratio of the height of the solid body tide to that of the deforming potential.

$k = 0.298$ for the solid Earth, or 0.245 for the Earth plus oceans, is the ratio of the additional tidal potential, produced by the re-distribution of mass, to the deforming potential.

$l \approx 0.084$ is the ratio of horizontal displacement of the crust to that of the equilibrium tide if the Earth were fluid.

These parameters are measures of the elasticity of the Earth as a whole. For a rigid Earth all three numbers are zero and for a fluid Earth in tidal equilibrium, $h_{\mathrm{f}} = 1$, $l_{\mathrm{f}} = 1$ and k_{f} is a function of the density profile,

$$k_{\mathrm{f}} \approx 3J_2/m \approx (2f_{\mathrm{H}}/m - 1), \tag{8.15}$$

where, in this equation, m is the ratio of the centrifugal component of gravity at the equator to the total. For a fluid Earth of uniform density $k_{\mathrm{f}} = 3/2$ (Problem 8.2) and for the actual density profile $k_{\mathrm{f}} = 0.937$.

Love numbers are used in analyses of several kinds of deformation and to distinguish the different values that are obtained, subscripts are often added. k_2 refers to a tidal deformation of ellipsoidal form represented by a degree 2 zonal harmonic. The value given above for the Earth plus oceans was determined from tidal perturbations of satellite orbits, as considered in the following section. A slightly different k, k_{W}, is obtained from the lengthening of the period of the Chandler wobble to $T_{\mathrm{W}} = 432$ days from the value, $T_{\mathrm{R}} = 305$ days, which applies to a hypothetical rigid Earth (Section 7.3). The gyroscopic torque (Eq. (7.22)) is reduced by deformation of the Earth, in partial accommodation of the misaligned equatorial bulge to the rotational potential. If the angular amplitude of the wobble, as seen in the latitude variation, is α, but the misalignment of the bulge is reduced to $(T_0/T_{\mathrm{W}})\alpha$, then the angular deformation of the mantle is $(1 - T_{\mathrm{R}}/T_{\mathrm{W}})\alpha$. For a fluid Earth the deformation would be α. Thus

$$\frac{k}{k_{\mathrm{f}}} = 1 - \frac{T_{\mathrm{R}}}{T_{\mathrm{W}}}. \tag{8.16}$$

However, we cannot directly identify k_{W} with k_2 because the oceans have quite different responses to the wobble and the semi-diurnal tide.

Earth-based measurements of tides do not give h or k directly, but only combinations of them. The marine tide responds to the total potential of the deformed Earth, but is observed relative to the deformed solid Earth. Similarly, tidal gravity is observed at sites which are themselves displaced by the tides. There is also a problem that, at the semi-diurnal and diurnal periods, the marine tides are far from equilibrium. There are longer-period components of the tides, that are presumed to be close to equilibrium, but they are small and are less easily observed. Satellite measurements, using a further development of the analysis in Chapter 9, give individual Love numbers directly and it is the satellite values that are the most precise, as well as the most reliable.

8.3 Tidal friction

The response of the Earth to the lunar tidal potential, W_2, is a deformation causing an additional potential k_2W_2. If the tides were perfectly linear and lossless, this would be an exact expression, but the tidal bulge is slightly delayed by turbulent drag in the sea and by anelasticity of the solid part of the Earth. Also the non-linear ocean response introduces higher harmonics. We can represent the delay as a departure of the prolate tidal elongation from alignment with the Earth–Moon axis by a small angle, δ. As discussed below, the measured value of δ is about 2.9°, so that the global high tide occurs at points that were directly in line with the Moon $(2.9/360) \times 24.84$ hours $= 12$ minutes ago. Thus, the tidal potential at the Earth's surface, due to the tidal deformation of the Earth itself, is

$$W_{\mathrm{E},a} = -\frac{k_2 G m a^2}{R^3}\left[\frac{3}{2}\cos^2(\psi - \delta) - \frac{1}{2}\right], \quad (8.17)$$

where m is the mass of the Moon, R is its distance, a is the Earth's radius and ψ is here the angle between the Earth–Moon axis and the radius to a surface point in the orbital plane. At any more remote point (r, ψ), that is at $r > a$, the potential, $W_{\mathrm{E},r}$, is reduced by the factor $(a/r)^3$ because the geometrical form of Eq. (8.17) is a second-degree zonal harmonic, similar to the second term of Eq. (6.15), and diminishes with distance r as r^{-3}.

$$W_{\mathrm{E},r} = -\frac{k_2 G m a^5}{R^3 r^3}\left(\frac{3}{2}\cos^2(\psi - \delta) - \frac{1}{2}\right). \quad (8.18)$$

Thus, there is a tidal torque exerted on any mass m^* at this point,

$$\begin{aligned} L_{\mathrm{T}} &= -m^* \frac{\partial W_{\mathrm{E},r}}{\partial(\psi - \delta)} \\ &= -\frac{3k_2 G m m^* a^5}{R^3 r^3}\cos(\psi - \delta)\sin(\psi - \delta). \end{aligned} \quad (8.19)$$

The observation of this torque by means of close satellites allows k_2 and δ to be determined (e.g. Christodoulidis *et al.*, 1988).

Now consider the tidal torque exerted on the Moon, for which $m^* = m$, $r = R$ and $\psi = 0$,

$$L_{\mathrm{T,Moon}} = \frac{3k_2 G\, m^2 a^5}{R^6}\cos\delta\sin\delta \approx \frac{3k_2\delta G m^2 a^5}{R^6}. \quad (8.20)$$

This torque acts in the direction that would reduce δ, that is, it tries to make the Moon 'catch up' with the tidal bulge of the Earth. The bulge appears to the Moon to be ahead of its own orbital motion, although with respect to the Earth, which is rotating faster, the bulge is seen to be delayed. Hence, the effect of the lag of the bulge, caused by frictional losses of the tides in the Earth, is to apply an accelerating torque to the orbital motion of the Moon. The equal torque exerted by the Moon on the bulge tends to pull the bulge into line with the Moon and so acts as a brake on the Earth's rotation (Fig. 8.4). Angular momentum is conserved. The angular momentum gained by the Moon is the same as that lost by the Earth because the torques are equal. However, energy is lost from the motion. The Moon's orbit gains energy (kinetic plus potential) at a rate $\omega_{\mathrm{L}} L_{\mathrm{T,Moon}}$ and the Earth's rotation loses energy at a greater rate $\omega L_{\mathrm{T,Moon}}$. The net rate of loss of energy is $(\omega - \omega_{\mathrm{L}})L_{\mathrm{T,Moon}} = 3.06 \times 10^{12}$ W, that is, the torque times the angular speed of the tide relative to the Earth. Adding the energy dissipation by the solar tide we have a total of 3.7×10^{12} W.

Observed values of the tidal parameters in Eq. (8.20) are $k_2 = 0.245$, $\delta = 2.89°$, giving a value

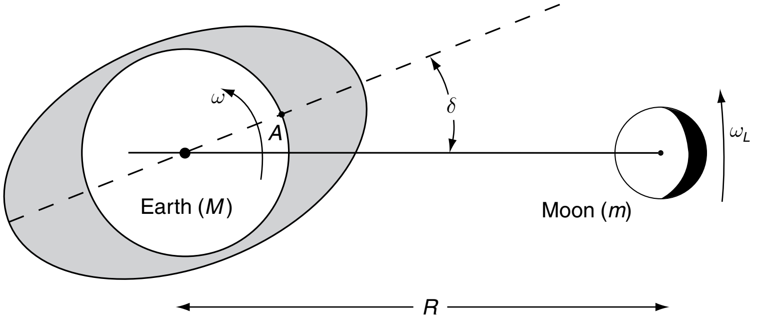


FIGURE 8.4 The origin of the tidal torque is a lag of the tidal bulge of the Earth relative to the Earth–Moon (or Earth–Sun) axis. The lag angle, δ, is about 3°. It exerts an accelerating torque on the Moon, causing its orbit to expand, and slows the Earth's rotation. The point A on the Earth was aligned with the Moon about 12 minutes ago, so the phase lag amounts to a delay of 12 minutes in the high tide.

of the lunar torque, $L_{\mathrm{T,Moon}} = 4.4 \times 10^{16}\,\mathrm{kg\,m^2\,s^{-2}}$. This is equated to the rate of increase in orbital angular momentum of the Earth–Moon system,

$$a_{\mathrm{L}} = \frac{mM}{m+M}\omega_{\mathrm{L}}R^2. \tag{8.21}$$

To identify separately the effects on ω_{L} and R, we use also Kepler's third law,

$$\omega_{\mathrm{L}}^2 R^3 = G(M+m), \tag{8.22}$$

from which

$$a_{\mathrm{L}} = \frac{G^{2/3}mM}{(m+M)^{1/3}}\omega_{\mathrm{L}}^{-1/3} \tag{8.23}$$

$$= \frac{G^{1/2}mM}{(m+M)^{1/2}}R^{1/2}, \tag{8.24}$$

so that

$$L_{\mathrm{T,Moon}} = \frac{\mathrm{d}a_{\mathrm{L}}}{\mathrm{d}t} = -\frac{1}{3}\frac{G^{2/3}mM}{(m+M)^{1/3}}\omega_{\mathrm{L}}^{-4/3}\frac{\mathrm{d}\omega_{\mathrm{L}}}{\mathrm{d}t} \tag{8.25}$$

$$= \frac{1}{2}\frac{G^{1/2}mM}{(m+M)^{1/2}}R^{-1/2}\frac{\mathrm{d}R}{\mathrm{d}t}. \tag{8.26}$$

These give

$$\frac{\mathrm{d}\omega_{\mathrm{L}}}{\mathrm{d}t} = -1.2 \times 10^{-23}\ \mathrm{rad\,s^{-2}}\left(25\,\mathrm{arcsec/century^2}\right), \tag{8.27}$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = 1.17 \times 10^{-9}\,\mathrm{m\,s^{-1}}(3.7\,\mathrm{cm/year}). \tag{8.28}$$

These results are derived from the satellite observations of tidal torque by Christodoulidis *et al.* (1988) who refer to lunar laser-ranging measurements by X. X. Newhall *et al.* that give a direct confirmation of Eq. (8.28). The decreasing angular speed of the Moon's orbital motion (Eq. 8.27) has been observed since the 1800s by conventional astronomy, but not with the precision that we now have with atomic clocks and satellite ranging.

The corresponding slowing of the Earth's rotation is given by

$$C\frac{\mathrm{d}\omega}{\mathrm{d}t} = -L_{\mathrm{T,Moon}}, \tag{8.29}$$

from which

$$\left(\frac{\mathrm{d}\omega}{\mathrm{d}t}\right)_{\text{lunar tide}} = -5.4 \times 10^{-22}\,\mathrm{rad\,s^{-2}}, \tag{8.30}$$

but this is not an independently observable quantity. Apart from the short term effects considered in Sections 6.4 and 7.4, there is a contribution to tidal braking of the Earth's rotation by the solar tide. As seen in Eq. (8.20), the tidal torque varies as $k_2\delta(m/R^3)^2$. For the Sun m^2/R^6 is smaller than for the Moon by the factor $(0.459)^2 = 0.21$, which is therefore the ratio of the contributions to $\mathrm{d}\omega/\mathrm{d}t$ if k_2 and δ are independent of tidal frequency. In general this is a doubtful assumption, but the angular frequencies of the present lunar and solar tides are sufficiently close that we cannot be far wrong by taking the present solar tidal braking to be 21% of the lunar effect. On this basis the total tidal braking of the Earth's rotation is

$$\left(\frac{\mathrm{d}\omega}{\mathrm{d}t}\right)_{\text{Total tide}} = -6.5 \times 10^{-22}\,\mathrm{rad\,s^{-2}}. \tag{8.31}$$

This corresponds to a length-of-day increase of 2.4 ms/century. However, at the present time there is partial cancellation by rotational acceleration caused by decreasing ellipticity, as discussed in Section 6.4.

The 2.9° phase angle is too large to explain in terms of dissipation in the solid Earth. It would require an anelastic *Q* factor of 1 / tan 2.9° = 20 for the mantle, which is many times lower than the observed *Q*, for which Ray *et al.* (2001) estimate a value of 280 at tidal periods (see also Section 10.5). Most of the tidal dissipation occurs in the oceans, where complexities of the geometry introduce local resonances and phase delays, especially in marginal seas and estuaries. But the satellite observations give us the global picture. The satellite value of k_2 is smaller than that calculated for the solid Earth from the seismically determined elasticity structure, demonstrating that the global average marine tide is inverted, that is, low tides appear where highs would be expected for an equilibrium tide. The driving speed of the tide, $(\omega - \omega_L)a = 450\ \mathrm{m\,s^{-1}}$ at the equator, is greater than the natural speed of the tidal wave, even in the deepest ocean: $v = \sqrt{gh} \approx 220\ \mathrm{m\,s^{-1}}$ in an ocean of depth $h = 5\,\mathrm{km}$. By driving the tide at a higher frequency than its natural resonance a large phase lag is developed. In the limit of zero dissipation the lag would be $\pi/2$. (Note that the tidal torque, Eq. (8.20) is zero if δ is any multiple of $\pi/2$).

The atmosphere is subject to a diurnal tide of thermal origin, as well as a gravitational tide. In spite of the tight coupling of atmospheric motion to the Earth, discussed in Section 26.2, the atmospheric tides are believed not to make significant contributions to the slowing rotation of the Earth and the thermal tide could accelerate it.

We can see from Eq. (8.14) that the tidal strain raised in one body by proximity to another one is proportional to the ratio of the masses. Therefore, the tidal deformation of the Moon by the Earth is about $(80)^2$ times the tide in the Earth. So, tidal friction in the Moon, if it were rotating, would be $(80)^4$ times as strong. Thus, we can see why the Moon's axial rotation coincides with its orbital period. Tidal friction has completely stopped its rotation relative to the Earth. The same is true for other close satellites in the Solar System. Io presents a constant face to Jupiter. In the case of Pluto and its large satellite, Charon, the relative rotations of both bodies have been stopped and they present fixed faces to one another.

8.4 Evolution of the lunar orbit

Equations (8.20) and (8.26) give a differential equation for the variation with time of the distance to the Moon,

$$R^{11/2}\frac{\mathrm{d}R}{\mathrm{d}t} = 6k_2\delta G^{1/2}a^5(m/M)(M+m)^{1/2}. \qquad (8.32)$$

If we make the simplest assumption, that $(k_2\delta)$ is constant, independent of the speed or amplitude of the tide, then the right-hand side of Eq. (8.32) is constant and we can write

$$R^{11/2}\mathrm{d}R/\mathrm{d}t = R_0^{11/2}R_0', \qquad (8.33)$$

where R_0 and R_0' are the present values of R and $\mathrm{d}R/\mathrm{d}t$. Then integrating from an initial distance, R_i, τ years ago,

$$\frac{2}{13}\left(R_0^{13/2} - R_\mathrm{i}^{13/2}\right) = R_0^{11/2}R_0'\tau. \qquad (8.34)$$

The high power of R ensures that, for R_i appreciably less than R_0, the early orbital evolution was very rapid and so the time scale hardly depends on the assumption made about R_i. The inferred total time for orbital evolution is

$$\tau \approx \frac{2}{13}R_0/R_0' = 1.6 \times 10^9\ \text{years}. \qquad (8.35)$$

The geology of the Moon has been stable for much longer than this and is incompatible with a close approach to the Earth as recently as 1.6×10^9 years ago. Also there is no evidence on the Earth for such a dramatic event. The assumption about $(k_2\delta)$ must be re-examined. Tidal friction was much weaker in the past than linear extrapolation from present conditions suggests.

The problem raised by Eq. (8.35) is emphasized by considering the asymptotic limit to which R tends, as one extrapolates backwards in time. This limit is given by the condition $\omega = \omega_L$, at which there would be no relative rotation and no moving tide or frictional loss. For the backward extrapolation it suffices to assume conservation of angular momentum of the

Earth–Moon system, neglecting the solar tide, but it is interesting to consider also the forward extrapolation, for which it is important to include the solar tide. Some assumption about the relationship between the solar and lunar tides is required and it is assumed here that $(k_2\delta)$ has the same value for both at all times, although it may vary arbitrarily with time. This assumption is likely to be seriously in error when the periods of the solar and lunar tides differ markedly, but it is better than the assumption that the solar tide is negligible.

Since the Earth–Sun distance changes insignificantly in the course of the orbital evolution of the Moon, the solar tide causes a retardation of the Earth's rotation,

$$\left(\frac{\mathrm{d}\omega}{\mathrm{d}t}\right)_{\mathrm{Solar}} = \left(\frac{\mathrm{d}\omega}{\mathrm{d}t}\right)_{\mathrm{present\ solar}} \times \frac{(k_2\delta)}{(k_2\delta)_{\mathrm{present}}}. \tag{8.36}$$

The present rate is $-1.2 \times 10^{-22}\,\mathrm{rad\,s^{-2}}$, being the difference between Eqs. (8.30) and (8.31). This frictional loss occurs independently of the Earth–Moon interaction, but by the assumption that $(k_2\delta)$ is the same for both, the two rates of tidal braking are linked. Consider a notional time, τ_n, which is the time required for a particular cumulative solar tidal effect if it had occurred at the present rate. Then by Eq. (8.34), τ_n is given in terms of the effect on the lunar orbit, so that the cumulative solar braking since the Moon was at distance R_i is

$$\begin{aligned}\Delta\omega_{\mathrm{solar}} &= \left(\frac{\mathrm{d}\omega}{\mathrm{d}t}\right)_{\mathrm{present\ solar}} \times \tau_n \\ &= \left(\frac{\mathrm{d}\omega}{\mathrm{d}t}\right)_{\mathrm{present\ solar}} \times \frac{2}{13}\frac{R_0}{R_0'}\left[1 - \left(\frac{R_i}{R_0}\right)^{13/2}\right].\end{aligned} \tag{8.37}$$

This is valid for arbitrary fluctuations of $(k_2\delta)$, as long as this product was the same for both solar and lunar tides. Validity of Eq. (8.37) does not require τ_n to be the actual time.

Now we can treat the lunar tide independently by conserving angular momentum, a_L, in the Earth–Moon system and writing it in terms of the present rotation rate, ω_0,

$$a_L = C\omega + \frac{Mm}{M+m}R^2\omega_L = 5.872C\omega_0, \tag{8.38}$$

so that, making ω the subject of this equation and adding $\Delta\omega_{\mathrm{solar}}$, we have

$$\omega = 5.872\omega_0 - \frac{Mm}{C(M+m)}R^2\omega_L - \Delta\omega_{\mathrm{solar}}. \tag{8.39}$$

Using this equation, we have a simple procedure for relating R, ω_L and ω at any stage of the orbital evolution, by selecting a value, R_i, of R, calculating the corresponding ω_L by Eq. (8.22) and then determining ω by Eq. (8.39) with $\Delta\omega_{\mathrm{solar}}$ given by Eq. (8.37). The result is shown in Fig. 8.5. The inner asymptotic limit, from which the Moon is receding, is found to be at $R = 2.3$ Earth radii, well inside the Roche limit of tidal stability, at which the Moon would have broken up (Section 8.5). There is no possibility that the Moon could ever have been so close.

The other asymptotic limit, at which $\omega = \omega_L$, is the one towards which the Moon is receding, at $R = 78$ Earth radii. After this point is reached, the solar tide will continue to slow the Earth's rotation, leaving the Moon orbiting faster than the Earth rotates. Then lunar tidal friction will start to cause the Moon to spiral back towards the Earth. However, unless $(k_2\delta)$ is then somehow dramatically enhanced, the time scale is so long that evolution of the Sun will overtake the Earth first.

Historical records of eclipses give evidence of the changes in ω and ω_L over 27 centuries, compared with the few decades of observations with atomic clocks. Since such observations give the drift in time, t, of the angular positions of the Earth and Moon (relative to the Sun) and tidal torques cause angular accelerations, the time drift accumulates as t^2 and so particular interest attaches to the earliest records. Following a demonstration by P. M. Muller that only reports of total eclipses were reliable, Stephenson and Morrison (1995) summarized the results of many studies of Chinese and Mesopotamian records back to 700 BC and concluded that they agree with the present value of $\mathrm{d}\omega_L/\mathrm{d}t$. However, $\mathrm{d}\omega/\mathrm{d}t$ has been more variable and there has been a persistent non-tidal acceleration of the rotation, attributed to post-glacial adjustment (Section 6.4).

On a geological time scale only the tidal effects are apparent. There have been several

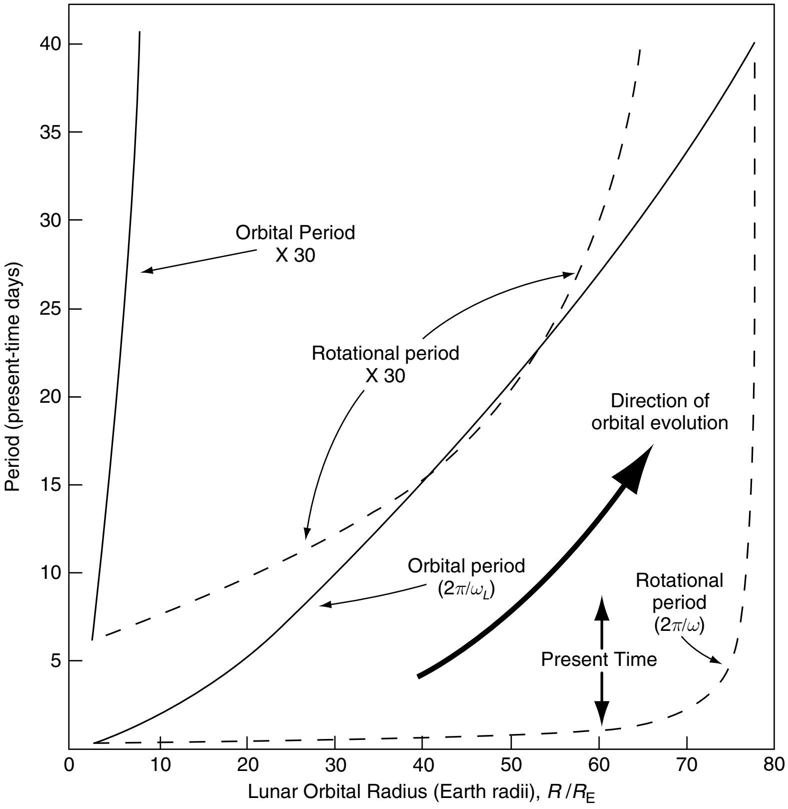


FIGURE 8.5 Variations in the orbital period of the Moon and the period of rotation of the Earth with the Moon's orbital radius, in the course of orbital evolution by tidal friction. Solar tides are allowed for by assuming that $(k_2\delta)$ (Eq. 8.20) is always the same for solar and lunar tides, although it may vary arbitrarily with time.

studies of 'paleontological clocks', the records of days per month and per year in the shells of marine organisms whose growth is controlled by daily, tidal and seasonal cycles. It is difficult to make reliable error estimates from such observations as there are various interruptions and irregularities to growth cycles, but the general conclusion is consistent. Rotation of the Earth and the lunar orbital motion were both faster some hundreds of millions of years ago, but not by as much as extrapolation of Eqs. (8.27) and (8.28) would suggest.

Reliable extensions of observations of tidal periods back to the early Precambrian period are obviously particularly valuable. This has been attempted using shells of marine creatures and stromatolites, calcareous mats deposited by communities of micro-organisms that are photosynthetic and grow towards the direction of the Sun, with daily growth increments and seasonal variations in growth direction. But the use of inorganic sedimentary markers controlled by tidal cycles, as observed by Williams (1990, 2000), has proved to be more satisfactory. We can compare Williams's observation of 400 ± 7 days per year 620 million years ago with the constant $(k_2\delta)$ extrapolation. This is the most reliable of the paleorotation measurements. Using Eq. (8.34) and Fig. 8.5, we obtain 443 days per year at that time. The average rate of slowing over the last 620 million years has been only about half that inferred from the present rate. Since Eq. (8.35) demonstrates that the average over the life of the Earth has been less still, there is apparently a trend to progressively increasing tidal dissipation, opposing the decrease due to the lunar recession. This is confirmed by the results of an investigation by Williams (2000) of banded rocks from Western Australia, dated at 2450 million years, indicating still weaker tidal friction in that early period (see Fig. 8.6).

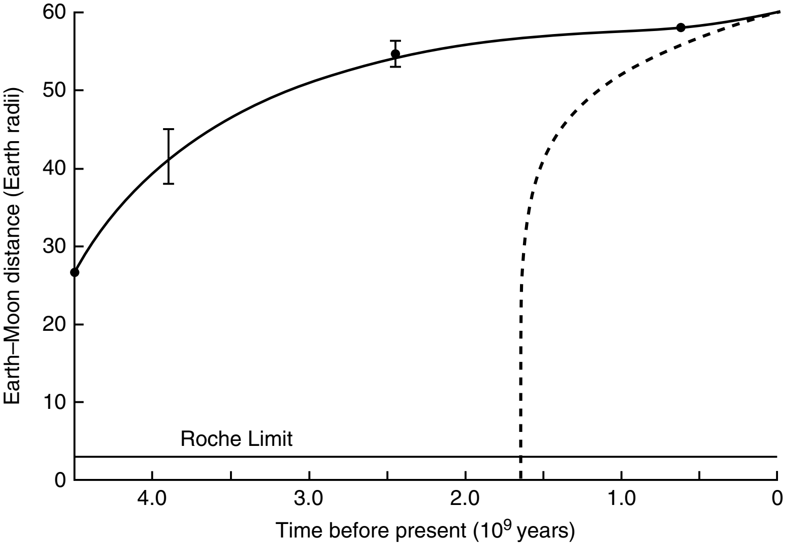


FIGURE 8.6 Earth–Moon distance as a function of time, using data from Williams (2000) at 620 my and 2450 my and the estimate at 3900 my from Section 8.6, compared with the extrapolation of present tidal friction by Eq. (8.34), shown as the broken line. This graph is very similar to Williams's Fig. 15, with the addition of the 3.9 Ga point, estimated from Eq. (8.56).

As mentioned, the marine tides are responsible for most of the tidal dissipation and they are locally very variable with large amplifications and phase delays, arising from complexities of sea floor geometry. It is sometimes presumed that present tidal friction happens to be strong as a fortuitous consequence of the present arrangement of the oceans and marginal seas. The progressive increase in continental crustal material (Section 5.3) and consequent raising of sea level to flood the continental margins may also be a contributory cause. While these hypotheses are admissible as plausible, we should not be too easily convinced by an explanation that requires the present oceans to be very different from those of all past periods. The consistent trend to decreasing effectiveness of tidal friction as one goes backwards in time accords better with the idea of Webb (1982) that tidal friction is increasing as the slowing tidal period approaches resonance with the free propagation of very long-period waves in the oceans.

Another factor that can be noted is that tidal dissipation occurs mainly by turbulence in marginal seas and estuaries, so that analysis as a linear phenomenon is questionable. Nonlinearity means that lunar and solar tides are not independent, but if we make the simple assumption that an arbitrary variation of dissipation with tidal amplitude can be represented by writing $\delta = \delta_0 (R/R_0)^k$ in Eq. (8.32), then Eq. (8.35) becomes

$$\tau = (R/R_0')/(13/2 - k). \qquad (8.40)$$

For any reasonable value of k this makes little difference to τ and the problem is not solved. We still seek a convincing explanation of the weak tidal friction of the early Earth, and comparison with a simple backward extrapolation of present day observations (Fig. 8.6) indicates that this is a serious shortcoming in our understanding of Earth history. The problem has prompted some exotic hypotheses for the origin of the Moon, involving a very close approach to the Earth, but this is not a feature of the lunar history presented in Section 8.6. We note also that the orbit evolution problem cannot be explained by a time variation of the gravitational constant, once considered possible, but securely dismissed (Hellings *et al.*, 1983).

8.5 The Roche limit for tidal stability of a satellite

Early in its history the Moon was closer to the Earth and we consider here the consequences of a very close approach. There is a limit imposed by the requirement for gravitational stability.

Although our view of lunar history does not include such a close approach, the analysis is directly relevant to a mutual approach by two moons (Section 8.6). The theory of tidal deformation, as presented in Section 8.2, assumes that it is slight. We now consider a situation in which that theory is inadequate: a very close approach to the Earth by a smaller body, such as the Moon. In this case we are interested in the tidal deformation of the smaller body by the Earth. Since the tide raised in the smaller body exceeds the tidal strain in the large one by the square of the ratio of the masses (Eq. 8.14), the integrity of the small one is endangered by the close approach. It becomes gravitationally unstable inside a critical separation, known as the Roche limit, after E. Roche, whose 1850 paper first considered the problem in detail.

Figure 8.7 represents the Moon (m) close to the Earth (M), with a strong tidal elongation in the direction of the Earth–Moon axis, identified as the c-axis of the Moon. The tidal elongation of the Earth along the same axis is sufficiently small to neglect in this calculation. The Moon is assumed to remain ellipsoidal under the extreme deformation; this is not quite correct but introduces no material error in the calculation of gravitational potential due to it. For points close to the Moon, and in particular on the surface itself, the approximations such as MacCullagh's formula (Eq. 6.7) are satisfactory only for slight ellipticities and we must use the general equation for the potential due to a prolate ellipsoid. This is given in convenient form by MacMillan (1958, p. 63). Referred to axes x, y, z, with the c axis aligned with z, the potential is

$$V = -\,G\pi\rho a^2 c\left(1+\frac{x^2+y^2-2z^2}{2(c^2-a^2)}\right)\frac{2}{\sqrt{c^2-a^2}}$$
$$\times\,\sinh^{-1}\left(\frac{c^2-a^2}{a^2+\kappa}\right)^{1/2}+\frac{G\pi\rho a^2 c\sqrt{c^2+\kappa}}{c^2-a^2}\cdot\frac{x^2+y^2}{a^2+\kappa}$$
$$-\,\frac{G\pi\rho a^2 c}{c^2-a^2}\cdot\frac{2z^2}{\sqrt{z^2+\kappa}},\qquad(8.41)$$

where κ satisfies the equation

$$\frac{x^2+y^2}{a^2+\kappa}+\frac{z^2}{c^2+\kappa}=1.$$

We are interested in the potentials at two particular points, P ($x=y=0, z=c$) and Q ($x^2+y^2=a^2; z=0$) in Fig. 8.7, at both of which $\kappa=0$ and Eq. (8.41) reduces to

$$V_{\mathrm{P}} = G\frac{2\pi\rho a^4 c}{(c^2-a^2)^{3/2}}\sinh^{-1}\left(\frac{c^2-a^2}{a^2}\right)^{1/2} - G\frac{2\pi\rho a^2c^2}{c^2-a^2},\qquad(8.42)$$

$$V_{\mathrm{Q}} = -\,G\frac{\pi\rho a^2 c\,(2c^2-a^2)}{(c^2-a^2)^{3/2}}\sinh^{-1}\left(\frac{c^2-a^2}{a^2}\right)^{1/2}$$
$$+\,G\frac{\pi\rho a^2c^2}{c^2-a^2}.\qquad(8.43)$$

The tidal ellipticity is determined by equating the total potentials at P and Q; thus to each of V_{P} and V_{Q} we add the gravitational potential due to the Earth (mass M) and rotational potentials of the two points about the centre of mass of the Earth–Moon system, assuming the Moon to keep a constant face towards the Earth, that is, the

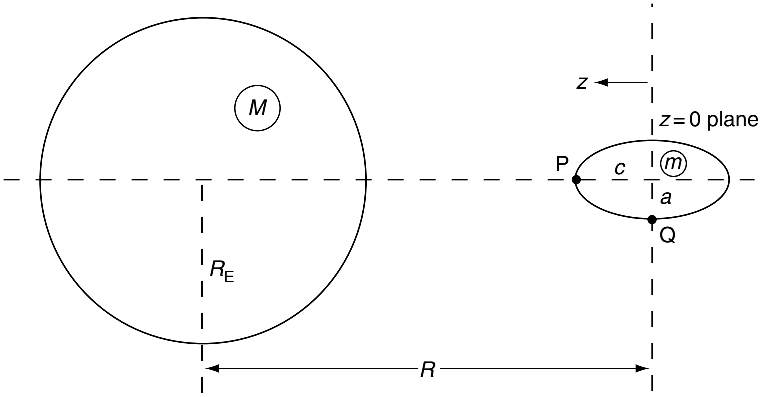


FIGURE 8.7 Geometry for the calculation of the critical distance R between a planet and its satellite for gravitational stability of the satellite.

axial rotation rate of the Moon is equal to its orbital rotation rate. Then

$$\begin{aligned} V_{\mathrm{P}} - \frac{GM}{R-c} - \frac{1}{2}\omega^2\left(\frac{M}{M+m}R - c\right)^2 \\ = V_{\mathrm{Q}} - \frac{GM}{R} - \frac{1}{2}\omega^2\left(\frac{M}{M+m}R\right)^2, \end{aligned} \tag{8.44}$$

where ω and R are related by Kepler's law (Eq. 8.22). Equation (8.44) simplifies to

$$\begin{aligned} V_{\mathrm{P}} - V_{\mathrm{Q}} &= \frac{3}{2}\frac{GMc^2}{R^3}\left[\frac{R\left(1+\frac{1}{3}\frac{m}{M}\right) - \frac{1}{3}c\left(1+\frac{m}{M}\right)}{R-c}\right] \\ &= \frac{3}{2}\frac{GM^*c^2}{R^3}, \end{aligned} \tag{8.45}$$

where we may regard

$$M^* = M\left[\frac{\left(1+\frac{1}{3}\frac{m}{M}\right) - \frac{1}{3}\frac{c}{R}\left(1+\frac{m}{M}\right)}{1-\frac{c}{R}}\right] \tag{8.46}$$

as an 'effective' mass of the Earth. M^* and M differ only to the extent that the mass of the Moon is not negligible with respect to that of the Earth and its semi-axis c with respect to the Earth–Moon distance, R.

From Eqs. (8.42) and (8.43) we have also

$$\begin{aligned} V_{\mathrm{P}} - V_{\mathrm{Q}} = &\frac{G\pi\rho a^2 c}{(c^2-a^2)^{3/2}}\left(2c^2+a^2\right)\sinh^{-1}\left(\frac{c^2-a^2}{a^2}\right)^{1/2} \\ &- \frac{G3\pi\rho a^2c^2}{c^2-a^2}, \end{aligned} \tag{8.47}$$

which may therefore be equated to Eq. (8.45). For this purpose it is convenient to represent the dimensions of the Moon in terms of the radius of a sphere of equal volume, $r_0 = (a^2c)^{1/3}$ and the ellipticity $e = \sqrt{(c^2-a^2)/a^2}$, so that

$$c = r_0\left(1+e^2\right)^{1/3},\ a = r_0\left(1+e^2\right)^{-1/6}.$$

Then, equating (8.45) and (8.47), we have

$$\begin{aligned} \frac{3}{2\pi\rho}\cdot\frac{M^*}{R^3} &= 2\frac{r_0^3}{m}\cdot\frac{M^*}{R^3} \\ &= \frac{1}{e^3}\left[\frac{3+2e^2}{\left(1+e^2\right)^{1/2}}\sinh^{-1}e - 3e\right]. \end{aligned} \tag{8.48}$$

(Expressing Eq. (8.47) in terms of r_0,e and expanding for small values of e, we obtain

$$V_{\mathrm{P}} - V_{\mathrm{Q}} = \frac{4\pi}{15}G\rho r_0^2 e^2 = \frac{1}{5}\frac{Gm}{r_0}e^2,$$

which coincides with the result obtained from the approximate Eq. (6.13) with the substitutions $C = 0.4ma^2$, $A = 0.2m(c^2+a^2)$ appropriate to a uniform ellipsoid.)

Now, at the Roche limit, the ellipticity is at the point of instability, that is, a small decrement in R causes the moon to break up or, in other words, its ellipticity grows indefinitely. Thus we can determine the Roche limit from the condition $\mathrm{d}e/\mathrm{d}R \to -\infty$ or $\mathrm{d}R/\mathrm{d}e \to 0$ applied to Eq. (8.48), which means that the derivative with respect to e of the right-hand side of Eq. (8.48) is equated to zero. This gives

$$\left(4e^4+14e^2+9\right)\sinh^{-1}e = \left(9e+8e^3\right)\left(1+e^2\right)^{1/2}, \tag{8.49}$$

the numerical solution of which is $e = 1.676$. Substituting this value into Eq. (8.48) gives

$$\frac{r_0^3}{m}\cdot\frac{M^*}{R^3} = 0.070\,31 \tag{8.50}$$

or

$$R = 2.42_3\left(\frac{M^*}{m}\right)^{1/3} r_0 = 2.42\left(\frac{M^*}{M}\right)^{1/3}\left(\frac{\rho_{\mathrm{E}}}{\rho}\right)^{1/3} R_{\mathrm{E}}, \tag{8.51}$$

where ρ_E, R_E are the density and radius of the Earth.

Ignoring for the moment the factor $(M^*/M)^{1/3}$, with $\rho_{\mathrm{E}} = 5515\,\mathrm{kg\,m^{-3}}$ and $\rho = 3340\,\mathrm{kg\,m^{-3}}$, Eq. (8.51) gives $R = 2.86_4 R_{\mathrm{E}}$. This allows an estimate of the term

$$\frac{c}{R} = \frac{c}{r_0}\cdot\frac{r_0}{R_{\mathrm{E}}}\cdot\frac{R_{\mathrm{E}}}{R} = 0.1435,$$

which is the principal contribution to the 'correction' factor $(M^*/M)^{1/3} = 1.037$, so that the 'final' result is

$$R = 2.97R_{\mathrm{E}}. \tag{8.52}$$

The correction for the finite mass and size of the Moon does not allow for the ellipticity of the Earth by virtue of the Moon's gravity, but the tidal ellipticity of the Earth is less than 1% of that of the Moon and even at $3R_{\mathrm{E}}$ its effect on the gravity field of the Earth at the Moon is negligible. We can also readily admit as satisfactory the assumption that the Moon is homogeneous; the moment of inertia of the Moon is sufficiently close to that of a body of uniform density that this cannot be a significant error. More difficult to assess is the assumption that the Moon remains ellipsoidal near to the Roche limit. Since, at the Roche limit, the point P in Fig. (8.7) is a neutral point in the field, the potential lines must cross there, so that although this figure properly represents the deformation of the Moon outside the Roche limit, at the limit itself the Moon must take the form of Fig. 8.8(a). With further decrease in separation the potential surface bounding the Moon opens, allowing material to escape, as in Fig. 8.8(b).

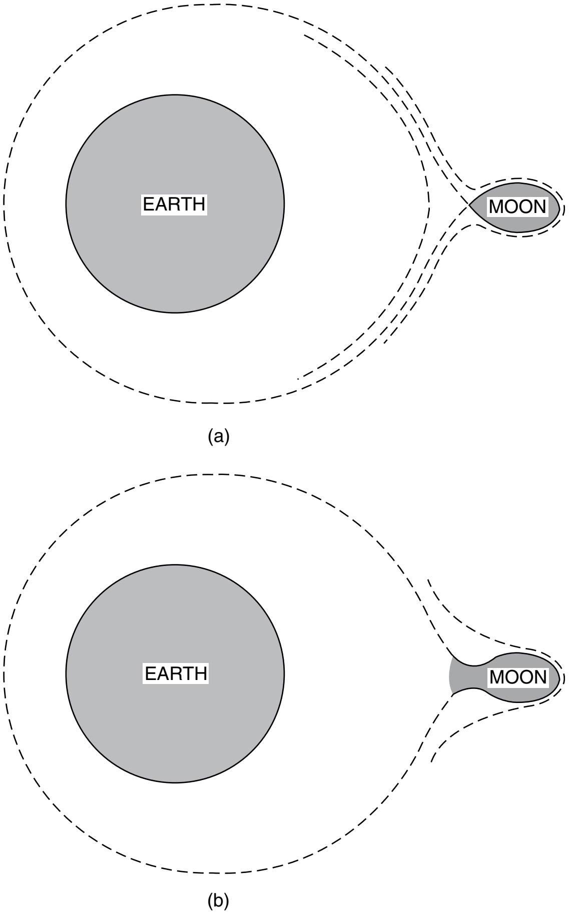


FIGURE 8.8 (a) Shape of the Moon at the Roche limit of gravitational stability. (b) Inside the Roche limit there is no bounding equipotential surface and the Moon breaks up. (Broken lines represent equipotential surfaces.)

8.6 The multiple moons hypothesis

Section 1.15 canvasses the idea that the Earth originally had at least two moons and that it is possible that Venus and Mercury also had satellites. The explanation for the disappearance of the additional bodies relies on tidal friction, and the equations in Sections 8.3 and 8.4 are used here to examine the idea more quantitatively. First, we note the operation of the Titius–Bode law in the accretion of the terrestrial planets and assume that the same principle applies to satellites. It is accurately obeyed by the major satellites of Jupiter (Section 1.2). We assume that that they remain independent as long as the ratio of their orbital radii exceeds the Bode's law ratio, about 1.6, but if they come within that separation then gravitational interaction will bring them together on the time scale of planetary accretion. This is much shorter than the interval, $\Delta\tau \approx 6 \times 10^8$ years, between formation of the Moon and its intense late bombardment, the lunar cataclysm discussed in Section 1.15. We examine the conditions required for tidal friction to bring two satellites together on this time scale.

We refer to our argument as the multiple moons hypothesis, and consider three moons, formed in orbits with radii differing by factors of about 1.6. The largest was probably innermost and subject to much stronger tidal friction than the others. Its orbit expanded, reducing its separation from the middle one, so that they quickly came closer than the Bode's law ratio, causing the smaller one to be subjected to orbit modification and eventual assimilation by the major one. That process was rapid enough to be observationally indistinguishable from the original planetary accretion, but the early existence of the 'middle' moon is necessary to explain the 600 million year delay before the major moon approached the outer one within the Bode's law orbital ratio. So, we calculate the time required for tidal friction to increase its orbital radius by the factor 1.6.

As in the discussion in Section 1.2, it is the ratio of orbital radii that we need to consider so we re-write Eq. (8.32) as a logarithmic derivative (and assume $m \ll M$):

$$\frac{\mathrm{d}\ln R}{\mathrm{d}t} = \frac{6G^{1/2}a^5}{M^{1/2}} k_2 \delta m R^{-13/2}. \tag{8.53}$$

This applies to each of two satellites with masses m_1 and m_2 at orbital radii R_1 and $R_2 > R_1$, so that

$$\frac{\mathrm{d}\ln(R_1/R_2)}{\mathrm{d}t} = \frac{6G^{1/2}a^5}{M^{1/2}} k_2 \delta \left(m_1 R_1^{-13/2} - m_2 R_2^{-13/2} \right). \tag{8.54}$$

The ratio R_1/R_2 increases with time, that is the orbits become closer on a logarithmic scale, if

$$m_1/m_2 > (R_1/R_2)^{13/2}. \tag{8.55}$$

If the satellites are to be brought together, then Eq. (8.55) must be satisfied at the critical ratio $R_1/R_2 = 1/1.6$, so that a necessary condition for them to merge is $m_1/m_2 > 0.047$. The inner one must have more than 5% of the mass of the outer one. This is an important restriction on the multiple moons hypothesis and is discussed below. Obviously the process occurs fastest if the inner satellite is more massive than the outer one and this is what we assume.

In considering the time scale we must allow that $(k_2\delta)$ was much smaller in the past than it is now. From Eq. (8.35) we know that the average value of δ over the life of the Earth was no more than a quarter of the present value (2.9°), but over the last 620 million years it averaged half of the present value (Section 8.4). It has increased with time and the value early in the history of the Earth was probably no more than 0.4°. If we make the extreme assumption that the Earth then had no ocean and rely on the anelastic Q of the solid Earth, but recognize that it was then hotter so that $Q \approx 100$ for mantle shear waves, then we would have $\delta \approx 0.2°$. This is probably too extreme for the Earth but appropriate for Venus and Mercury, which lack oceans. Thus, by assuming $\delta = 0.4°$, we can hardly be wrong by a factor exceeding 2.

As in the Roche limit analysis in Section 8.5, when two bodies approach one another within a critical distance the smaller one becomes gravitationally unstable and, under appropriate conditions, which include modest relative speed, breaks up. The fact that the Moon has preserved a record of the cratering without disruption of its sphericity is evidence that it was much larger than any of its impactors and that the sum of them, that is the second satellite, was probably below the 5% mass ratio referred to above. In that case it is necessary to suppose that the smaller satellites were initially more remote from the Earth. It also means that we can neglect the tidal evolution of their orbits and estimate the time scale of the interaction from the behaviour of the larger body, which approximated the present Moon. This is calculable by extrapolation from the present orbital evolution with the adjustment for δ mentioned above.

Integrating Eq. (8.33) from an initial radius R_1 to R_1^* over $\Delta\tau = 6 \times 10^8$ years, with R_0' (3.7 cm/year at the present time) reduced by the factor (0.4/2.9) to account for the lower value of δ, we have

$$\begin{aligned}\left(\frac{R_1^*}{R_0}\right)^{13/2} - \left(\frac{R_1}{R_0}\right)^{13/2} &= \frac{13}{2}\frac{0.4}{2.9}\frac{R_0'}{R_0}\Delta\tau \\ &= 0.05. \end{aligned} \tag{8.56}$$

For any value of R_1/R_0 less than about 0.4, $R_1^*/R_0 = 0.63$ and R_2/R_0 is close to 1.0, placing the smaller body almost in the present lunar orbit. When the impacts occurred, about 3.9 billion years ago, the Moon was in an orbit of at least 38 Earth radii and possibly 45 Earth radii (compared with the present 60.3 Earth radii). This is supported by the Earth–Moon distance 2450 million years ago, as estimated by Williams (2000), ~55 Earth radii, and is consistent with weak early tidal friction, as discussed in Section 8.4.

Now we can examine the consequences of tidal friction in the first 600 million years. The assumption that the orbital radii of the satellites were initially separated by the Bode's law ratio, 1.6, means that this is the factor by which the Moon's orbit evolved over that 600 million year period. The rate of orbital evolution is a very strong function of orbital radius, so that, if the phase lag, δ, is specified, the radii at the beginning and end of that period are fixed. Assuming $\delta = 0.4^\circ$, as before, the Moon started at $25R_E$ and reached $40R_E$ at the time of the cataclysm, 600 million years later, in agreement with the extrapolation of the Williams (2000) data in Fig. 8.6. This argument imposes a strong constraint on the distance at which the Moon formed. If that had been much less than $25R_E$, then the initial orbital radius would have increased very rapidly and grown by the factor 1.6 in much less than 600 million years. No plausible value of δ would have made that process last for 600 million years. An initial lunar distance of $25R_E$ is incompatible with the giant impact hypothesis for the origin of the Moon, according to which it would have formed inside about $4R_E$. These numbers all fit the history of the lunar orbit as plotted in Fig. 8.6, with the Moon migrating from 26 Earth radii to 40 Earth radii (the factor 1.6) in its first 600 million years and thereafter more slowly. We echo the words of Williams: 'This tidal scenario suggests that a close approach of the Moon has not occurred at any time in Earth history'.

The details of the calculation are surprisingly well constrained. It demonstrates that what had been a paradoxical observation, the late intense bombardment of the Moon, can be explained by familiar physical processes without requiring any bombardment of the Earth. The fragments that impacted the Moon were in Earth orbit, as was the body from which they were derived. This body was 'stored' in a terrestrial orbit for 6×10^8 years of relative lunar quiescence before the orbit was disturbed by the approaching Moon. It is important that both bodies were in terrestrial orbits, because this means that the 'fatal' approach was slow enough to break the small one into several pieces, which probably made further disrupting approaches before impacting. The evolving orbit of the major moon would not have been dramatically affected by interaction with the smaller ones.

While our orbital history of the Moon is obviously conjectural and is subject to uncertainties in some of the numbers we have used, it is based on the well understood principle of tidal friction to address two paradoxes: the lunar cataclysm 3.9 Ga ago and the fewness of satellites in the inner Solar System. At the same time we avoid what we see, in Section 1.15, as difficulties with the giant impact hypothesis for the origin of the Moon.

9

The satellite geoid, isostasy, post-glacial rebound and mantle viscosity

9.1 Preamble

Gravity observations are referred to an equipotential surface, termed the geoid, for which sea level is a close approximation. We can picture the geoidal surface in continental areas as following the water level in hypothetical narrow canals connected to the oceans. For a non-rotating planet in hydrostatic equilibrium the geoid would be a sphere but rotation deforms it to an oblate ellipsoid. For several reasons discussed in Section 6.4 the Earth is slightly more elliptical than equilibrium theory would suggest. One reason is the depression of polar regions by former glaciation, from which recovery is incomplete and the continuing rebound gives a clue to the viscosity of the mantle. There are heterogeneities at all levels, the most obvious of which are seen at the surface as continents and oceans, but the effect of the continent–ocean structure on the geoid is barely discernible and very much less than if the continents were superimposed on an otherwise uniform Earth. On a continental scale the surface features are very nearly in hydrostatic balance. This is the principle of isostasy.

Features of the gravity field on a scale larger than 1000 km are discerned more effectively by studying perturbations of satellite orbits than would be possible from surface observations. With progressive improvements, satellite techniques have been used to distinguish increasingly fine features, although for exploration of local anomalies surface observations on land must still be used, sometimes in combination with satellite data. The shape of the sea surface is observed by satellite altimetry, radar reflections to satellites. It follows the geoid quite closely and shows finer scale details than could be inferred from orbital motion.

Gravitational perturbations of satellite orbits are analysed in terms of spherical harmonics (Appendix C). In Section 6.2 we consider just the centred mass and ellipticity. The effect of ellipticity is smaller than the central gravitational attraction by a factor of 1000 for satellites in low orbits, $r \approx a$. The higher terms in a more general harmonic expansion, representing finer details of the field that we now consider, are smaller still, by another factor $\sim$1000. For this purpose the harmonic terms are fully normalized, as defined by Eqs. (C.13) and (C.14) in Appendix C, so that the mean square values over the surface of a sphere are unity. This means that the coefficients $\bar{C}_l^m$and $\bar{S}_l^m$, representing departures from sphericity, relate directly to the amplitudes of the features represented. Referred to fully normalized harmonics the ellipticity coefficient is not J_2 (Eq. (6.14)), but $\bar{C}_2^0 = -J_2/\sqrt{5}$. However, if interest is restricted to ellipticity, then J_2 is used.

For more than two decades, satellite measurements of J_2 have been precise enough to observe a slow decrease with time, $\dot{J}_2 = -2.8 \times 10^{-11}$ per year. The rate is clearly greater than can be explained by tidal braking of the Earth's rotation (Section 8.4) and is attributed to post-glacial rebound, although the adequacy of this to explain the total $\dot{J}_2$ remains to be confirmed. The transfer of mass into the depressed areas from lower latitudes causes a decrease in the axial moment of inertia, C, and this is apparent from the corresponding decrease in J_2. Unlike tidal friction, which transfers rotational angular momentum to the lunar and solar orbits, rebound conserves rotational angular momentum and so causes spin-up, partially offsetting the rotational slowing by tidal friction (Section 6.4).

As mentioned in Section 6.4, $\dot{J}_2$ cannot be used directly to infer relaxation of excess J_2, because the excess attributable to glaciation cannot be separated from other, unrelated components. Instead $\dot{J}_2$ is interpreted as a response to a known history of ice loading. Of the areas of former deep glaciation, where rebound continues at an observable rate, the two that have been studied in closest detail are Fennoscandia, centred on the Gulf of Bothnia, and Laurentia, centred on Canada. We concentrate attention on Laurentia, which is much bigger and, according to Peltier's (2004) ice model, accounts for as much as 2/3 of the global deglaciation. Antarctica ranks second, although it still holds more ice than it is estimated to have lost. We use Laurentia to estimate the fraction of $\dot{J}_2$ that is explained by it, with the conclusion is that there is a bigger ellipticity effect than is explained by the documented areas of rebound.

From a fundamental perspective the most interesting conclusion of rebound studies is the viscosity of the mantle and its depth dependence. These studies assume that the mantle rheology is linear, as for a Newtonian viscous fluid. This assumption is insecure, and the possibility of non-linear rheology is considered in Section 10.6. Non-linearity means that different values of effective viscosity would apply to processes with different strain rates. In applying the rebound estimate of viscosity to mantle convection in Section 13.2, we conclude that these two phenomena are consistent with similar values of viscosity. However, the strain rates for plate motion and post-glacial rebound are also similar. The same value of effective viscosity would be expected for both, whether or not the mantle theology is linear. This gives confidence that both convection and rebound are satisfactorily explained, but the question of linearity remains unanswered.

9.2 The satellite geoid

To a first approximation the effect of the Earth's gravity on a satellite is given by the first term of Eq. (6.2) or (6.13), which describes the force maintaining the satellite on an elliptical orbit (Appendix B). The following discussion considers first the effect of the second term in these equations and then higher order terms. As considered in Section 7.2, the second term causes a mutual precessional torque between the Earth and the satellite. For a man-made satellite the effect on the Earth is negligible, but the satellite orbit precesses at a rate that provides a measure of the ellipticity coefficient J_2 (Eq. (6.14)). This is referred to as a regression of the nodes of the orbit, that is a progressive drift of the points at which it crosses the orbital plane (as seen by an observer fixed in space, not rotating with the Earth). Equations describing this process follow closely those used in Section 7.2 for precession of the Earth.

Considering a satellite in an orbit of radius r, inclined at an angle i to the equatorial plane, we may rewrite Eq. (7.10) to give the mean torque over a complete orbit, acting on the satellite,

$$\bar{L} = -\frac{3}{2}\frac{Gm}{r^3}Ma^2J_2\cos i\sin i, \tag{9.1}$$

where m is the satellite mass and M is the Earth's mass. This torque acts on the component of the orbital angular momentum that is perpendicular to the Earth's axis,

$$p_\perp = mr^2\omega_S\sin i, \tag{9.2}$$

where ω_S is the orbital angular speed of the satellite. The mean angular rate of the orbital precession, observed as a steady regression of the nodes of the orbit, that is the positions in space where they cross the equatorial plane, is the ratio of the torque to the angular momentum component on which it acts,

$$\overline{\omega}_P = \frac{\overline{L}}{p_\perp} = -\frac{3}{2}\frac{GMa^2J_2}{r^5\omega_S}\cos i. \qquad (9.3)$$

This simplifies by applying Kepler's third law $\left(\omega_S^2 r^3 = GM\right)$ to the orbit:

$$\overline{\omega}_P = -\frac{3}{2}\omega_S\frac{a^2}{r^2}J_2\cos i. \qquad (9.4)$$

Thus the angular change per orbital revolution in the position of a node is $\Delta\Omega$, given by

$$\frac{\Delta\Omega}{2\pi} = \frac{\overline{\omega}_P}{\omega_S} = -\frac{3}{2}\frac{a^2}{r^2}J_2\cos i. \qquad (9.5)$$

For a close orbit (at $r = 1.03\ a$) inclined at $i = 45°$, $\Omega = 0.38°$.

Equation (9.5), or a more general version for an elliptical orbit, suffices for a determination of J_2 to an accuracy of order 1 part in 1000, but for greater accuracy higher harmonics in the gravitational potential must be taken into account. Variations with longitude (tesseral harmonics, as given by Eq. (C.15) in Appendix C, with $m \neq 0$) are averaged out by taking a long record of nodal regression, but all zonal harmonics (with $m = 0$ and no longitude dependence) contribute to the long-term average of $\Delta\Omega$. The contributions depend in different ways on the inclinations and radii of the orbits, so that data from several satellites are required to separate them. Detailed observations of elliptical orbits also provide additional information.

In applying Eq. (C.15) to the general problem of the geoid, only the unprimed coefficients, representing sources of internal origin, are of interest. It is convenient also to write the $l = 0$ term, representing the central mass, separately, and to note that $l = 1$ gives an asymmetrical potential and that it vanishes by the selection of the centre of mass of the Earth as the coordinate origin. Then, by normalizing the sum of all terms to (GM/r), the potential in standard form, with fully normalized coefficients (Eq. (C.13)), is

$$V = -\frac{GM}{r}\left\{1 + \sum_{l=2}^{\infty}\left(\frac{a}{r}\right)^l \sum_{m=0}^{l} \mathrm{p}_l^m(\sin\phi) \left[\overline{C}_l^m\cos m\lambda + \overline{S}_l^m\sin m\lambda\right]\right\}. \qquad (9.6)$$

Care is required with signs, as well as normalization, in the application of this equation. Thus, for example, $\overline{C}_2^0 = -J_2/\sqrt{5}$, as mentioned in Section 9.1.

Tesseral harmonics (with $m \neq 0$) in the gravitational potential cause shorter-period variations in satellite orbits (such as an oscillation in the rate of nodal regression) than do the zonal harmonics. More detailed observations are required to measure them. It follows also that they are less accurately determined. Nevertheless, there have been numerous independent studies and they have converged to good agreement for harmonics to $l \approx 50$. Values of geoid coefficients to degree and order (8,8) are listed in Table 9.1, and Fig. 9.1 is a map of geoidal elevation obtained from coefficients to degree and order (50,50). Rapp and Pavlis (1990) reported a calculation of geoid coefficients to harmonic degree 360 with a full geoid map.

The spherical harmonic representation of the geoid is particularly appropriate for analysis of its broad features and follows naturally from the study of perturbations in satellite orbits, which provide the most accurate measure of the low-degree terms. Spherical harmonic expansions become cumbersome when extended to high harmonic degrees, l, because, for $l \geq 2$, the number of coefficients varies as $(l^2 + 2l - 5)$. There is also a problem of uneven coverage of the Earth when satellite orbit data are supplemented by surface measurements to fill in the fine details. However, satellite altimetry, the timing of radar reflections to satellites from the sea surface, has given considerable detail in

Table 9.1 Spherical harmonic coefficients of the Earth's gravitational potential to degree and order (8,8), from a more extensive set by Lerch *et al.* (1994). For each (l,m) the coefficients are $\bar{C}_l^m$ followed by $\bar{S}_l^m$, in units of 10^{-6}, as in Eq. (9.6) and defined in Appendix C

$l\backslash m$	0	1	2	3	4	5	6	7	8
2	–	–	2.439	–	–	–	–	–	–
	484.165		– 1.400						
3	0.957	2.029	0.904	0.720	–	–	–	–	–
		0.249	−0.619	1.414					
4	0.539	−0.536	0.349	0.991	−0.188	–	–	–	–
		−0.473	0.664	−0.201	0.309				
5	0.069	−0.061	0.655	−0.452	0.296	0.175	–	–	–
		−0.096	−0.325	−0.217	0.050	−0.668			
6	−0.148	−0.076	0.052	0.057	−0.088	−0.267	0.010	–	–
		0.026	−0.376	0.009	−0.472	−0.536	−0.237		
7	0.090	0.280	0.323	0.251	−0.275	0.002	−0.359	0.001	–
		0.096	0.096	−0.212	−0.128	0.019	0.152	0.024	
8	0.047	0.023	0.073	−0.018	−0.244	−0.025	−0.065	0.069	0.123
		0.060	0.069	−0.087	0.068	0.088	0.309	0.076	0.122

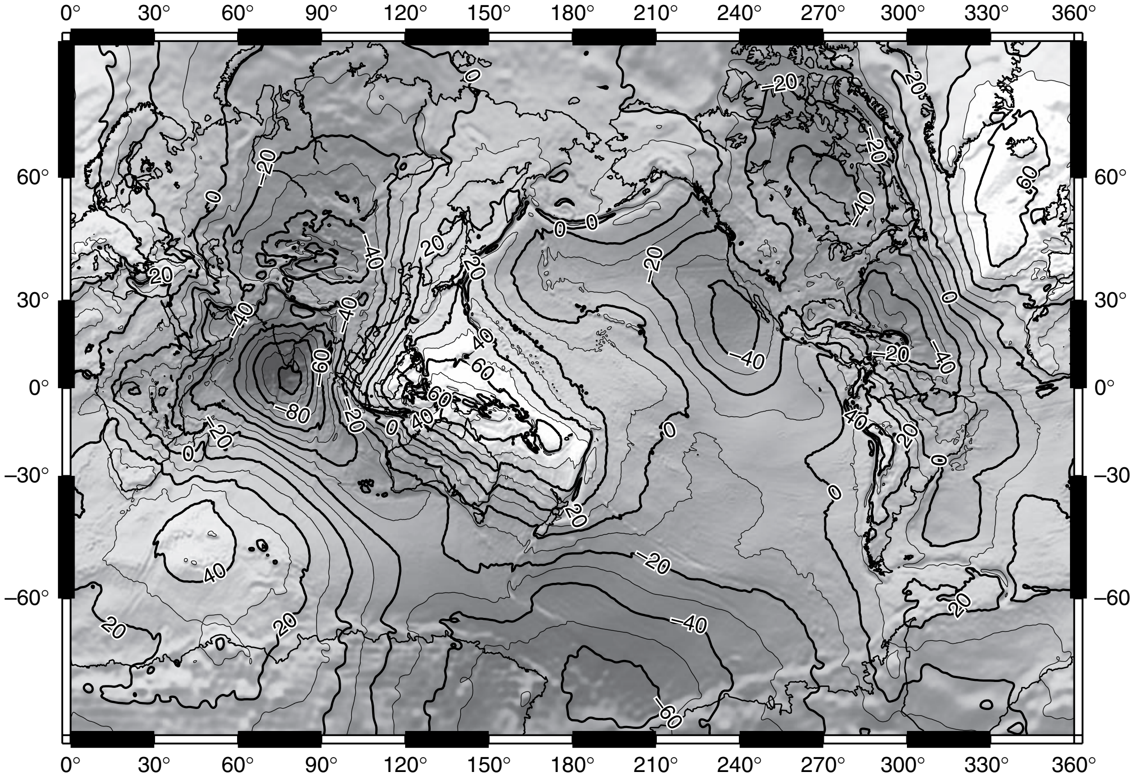


FIGURE 9.1 Contours of geoid height (in metres), relative to the reference ellipsoid (f = 1/298.257). Courtesy of David Sandwell.

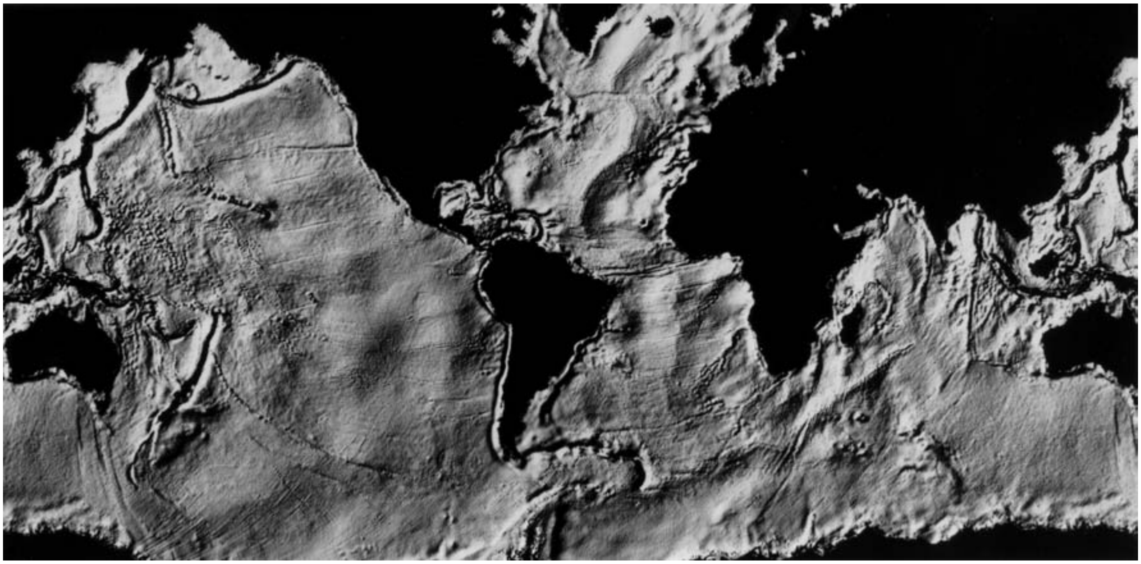


FIGURE 9.2 The mean sea level, as mapped by satellite altimetry, reflects the geoid in oceanic areas. Reproduced from Cazenave (1995).

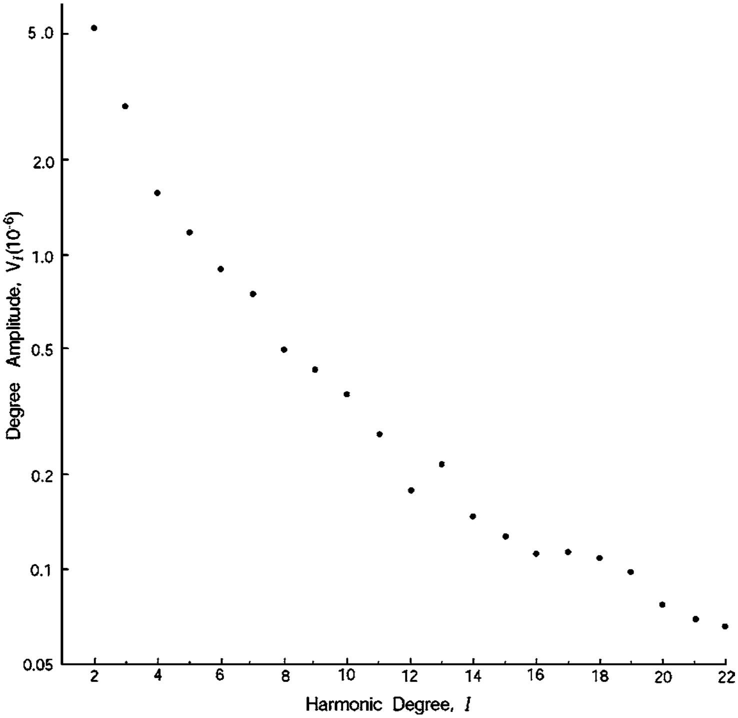


FIGURE 9.3 Degree amplitudes, as defined by Eq. (9.7), of the spherical harmonic coefficients of the geoid listed by Lerch *et al.* (1994). Values corresponding to the equilibrium ellipticity have been subtracted from $\bar{C}_2^0$ and $\bar{C}_4^0$.

oceanic areas. It demonstrates the close correspondence of the fine structure of the geoid to tectonic features of the Earth (Fig. 9.2), as reflected in the sea floor topography, which the sea-surface geoid mimics.

The relationship of the low-degree harmonic terms to the internal structure and dynamics of the Earth is less obvious. The spectrum of harmonic degree amplitudes plotted in Fig. 9.3 indicates that the low degrees represent deeper features than those responsible for the details seen in Fig. 9.2. Since spherical harmonics are orthogonal functions, the total amplitude of all harmonic terms of degree l, taken together, is

$$V_l = \left\{ \sum_{m=0}^{l} \left[\left(\bar{C}_l^m \right)^2 + \left(\bar{S}_l^m \right)^2 \right] \right\}^{1/2}, \qquad (9.7)$$

and this is the function plotted in Fig. 9.3. V_l is the rms amplitude (square root of spectral power) of geoid undulations with wavelength $(2\pi R_E/l)$, where R_E is the Earth's radius. The spectrum in Fig. 9.3 is very 'red', that is amplitude increases with wavelength, or l^{-1}. This is indicative of deep mantle sources. The approximately linear decrease in $\log V_l$ with l over the range $l = 4$ to 16 is suggestive of a spectrally 'white' source at a radius at which $d(\ln V_l)/dl$ would be zero. This would put it well into the lower mantle. Although a source at a single depth could not be physically realistic, and density heterogeneities are widely distributed, the broad-scale features of the satellite geoid indicate that the heterogeneities extend deep into the lower mantle.

If such an analysis is extended to higher harmonic degrees, then it is obvious that the spectral 'redness' cannot persist indefinitely. Deep-seated contributions to high-degree components of the geoid cannot be seen at the surface and effects of shallower sources become dominant.

9.3 The principle of isostasy

The distribution of elevations of the Earth's solid surface is strongly bimodal, as illustrated in Fig. 9.4. Most of the surface is either of continental type, with an elevation within 1 km of sea level, or of oceanic type, 4 to 5 km below sea level. Taking the average density difference between the continental crust and sea water to be 1750 kg m^{-3}, the continents have a mass excess, relative to the oceans, of 8×10^6 kg m^{-2} above the level of the ocean floor. Such a mass excess is not evident in the broad scale features of the geoid in Fig. 9.1. Thus this mass is compensated by continental 'roots' of material with lower density than that of the sub-oceanic mantle. This is the principle of isostasy, first stated in the mid-nineteenth century when it was recognized that the gravitational deflection of the vertical by the Himalaya mountains was much less than if they were simply a protrusion on an otherwise spherically layered Earth. As we now know from seismic studies, the continental crust is typically 35 km to 40 km thick (60 to 75 km under the Himalaya), whereas the oceanic crust is only 7 km thick, and both overlie the denser mantle. We consider a plausible reason for the continental thickness at the end of this section, and, in Section 23.2 make use of the idea that it has been constant as the crust developed by differentiation from the mantle.

For an idea of the effect of continents on the geoid, consider an idealized pair of circular continents on a spherical Earth, as in Fig. 9.5. This simple geometry allows their broad scale (ellipsoidal) deformation of the geoid to be calculated from their moments of inertia by Eq. (6.13). In case (a) the moments of inertia of the Earth about axes 1 and 2 differ only by the moments of inertia I_1 and I_2 of the continents about these axes. In the approximation that the radii, r, of the continents, as well as their thicknesses, h, are much less than the radius of the Earth, $h \ll r \ll a$,

$$I_1 \approx 2(\pi r^2 h \rho_c) r^2/2, \qquad (9.8)$$

$$I_2 \approx 2 \int_a^{a+h} (\pi r^2 \rho_c) x^2 dx \approx 2\pi r^2 \rho_c a^2 h, \qquad (9.9)$$

so that

$$I_1 - I_2 \approx -2\pi r^2 \rho_c a^2 h. \qquad (9.10)$$

Thus

$$J_2 = \frac{I_1 - I_2}{Ma^2} \approx -2\pi \rho_c r^2 h/M, \qquad (9.11)$$

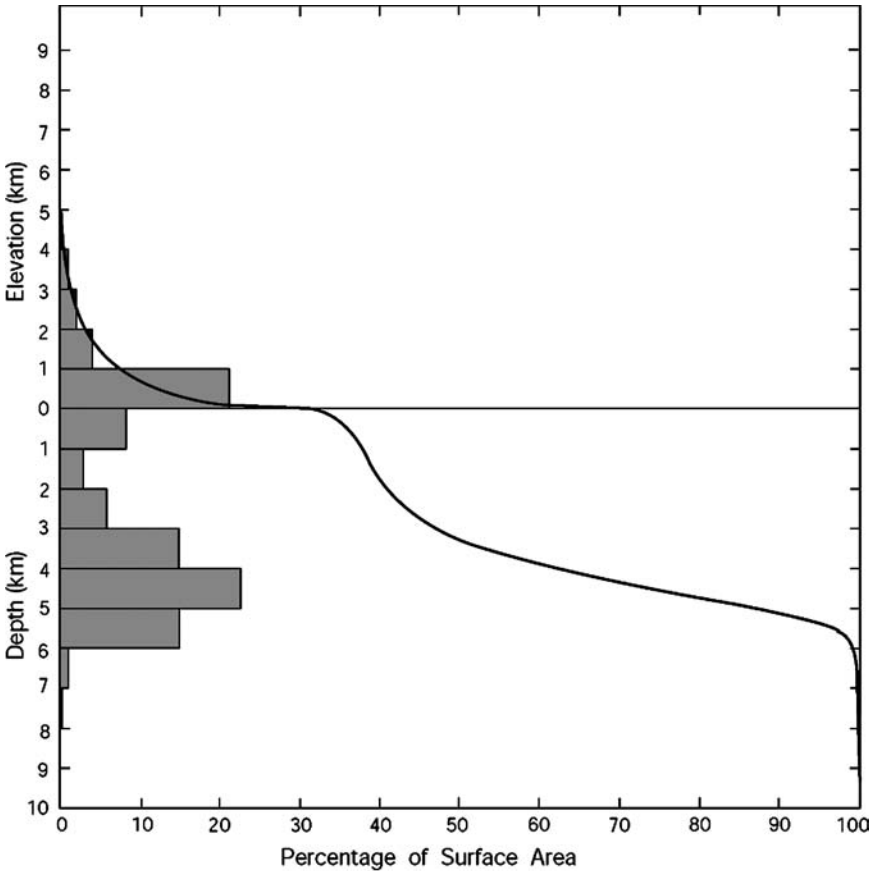


FIGURE 9.4 The hypsographic curve – the characteristic distribution of the elevations of the solid surface – with a histogram of areas in 1000 m elevation intervals.

and by Eq. (6.20), with rotation not considered,

$$f = \frac{3}{2}J_2 \approx -3\pi\rho_c r^2 h/M \approx -\frac{9}{4}\frac{r^2 h}{a^3}\frac{\rho_c}{\bar{\rho}}, \tag{9.12}$$

where $\bar{\rho}$ is the mean density of the Earth. The negative flattening means a prolate elongation along axis 1. The geoid elevation, h_g, on this axis, relative to axis 2, is a fraction of the continental elevation, h, given by

$$\frac{h_g}{h} = -\frac{fa}{h} = \frac{9}{4}\frac{r^2}{a^2}\frac{\rho_c}{\bar{\rho}}. \tag{9.13}$$

Taking $r = 2500$ km as representative of a continent, and $\rho_c = 1750$ kg m^{-3} as the density difference between the continent and sea water,

$$h_g \approx 0.11h. \tag{9.14}$$

If this analysis were applicable to the real Earth, we would see systematic differences in geoid heights between centres of continents and oceans of order 0.11×4500 m $= 500$ m. Not only are geoid height variations very much smaller than this, but the variations that are observed (Fig. 9.1) are not obviously related to the continent–ocean structure. On a continental scale isostatic balance is closely maintained, although rigidity of the lithosphere supports isostatic anomalies on a scale of order 100 km.

Now consider the continental model in Fig. 9.5(b) and require that it cause no geoid ellipticity. This requirement is satisfied by making the moments of inertia of the continents equal to that of the oceanic crust and mantle that would replace them to produce a spherically symmetrical Earth:

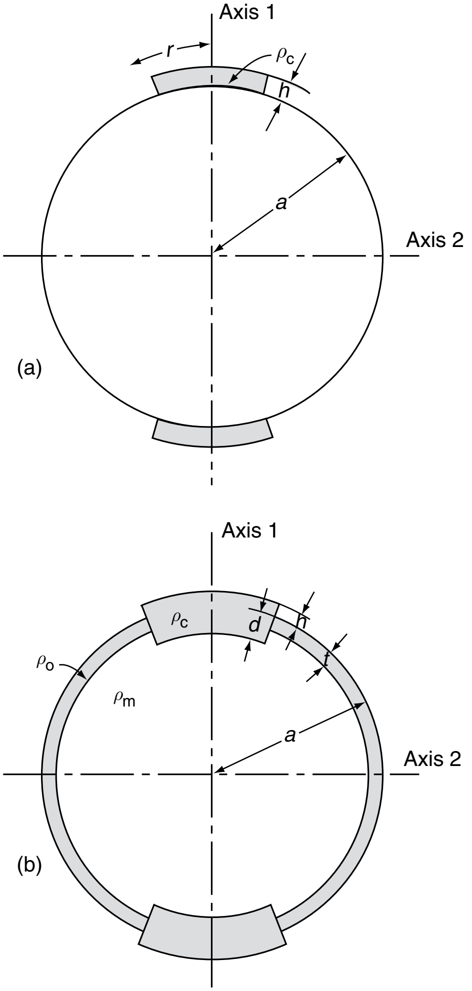


FIGURE 9.5 Simple models of a symmetrical pair of continents, illustrating their effect on the geoid. (a) Continents superimposed on an otherwise spherically symmetric Earth. (b) Continents of density ρ_c overlying a mantle of density ρ_m and isostatically balanced with an oceanic crust of density ρ_o.

$$2\int_{a-d}^{a+h}(\pi r^2\rho_c)x^2\mathrm{d}x = 2\int_{a-d}^{a-t}(\pi r^2\rho_m)x^2\mathrm{d}x + 2\int_{a-t}^{a}(\pi r^2\rho_o)x^2\mathrm{d}x. \qquad (9.15)$$

With the conditions $d, t, h \ll a$, this gives

$$\rho_c(h+d) = \rho_m(d-t) + \rho_o t. \qquad (9.16)$$

which is a statement of the principle of isostasy, that the total masses in all vertical columns (of unit area) are the same.

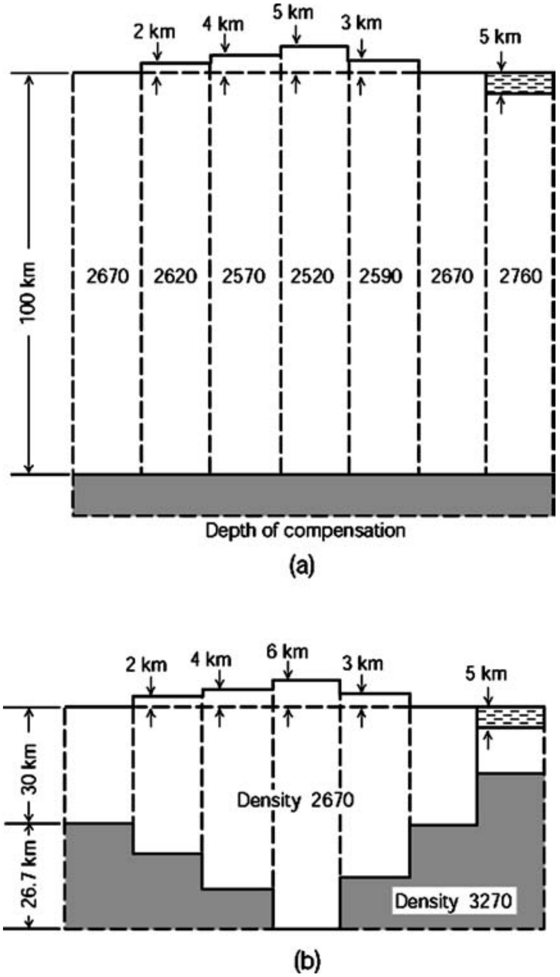


FIGURE 9.6 Isostatic compensation according to (a) J. H. Pratt and (b) G. B. Airy, with numerical values of density by W. A. Heiskanen. Continents, ocean basins and mountain ranges are balanced by either of these principles. Crustal structure corresponds more nearly to (b).

Equation (9.16) incorporates both of the rival hypotheses that were inspired by the evidence that the Himalaya are isostatically balanced. They express alternative methods by which isostasy may be achieved, as in Fig. 9.6. In 1854 J. H. Pratt suggested that the higher parts of the crust were elevated by virtue of their lower densities, with a common compensation depth, as in Fig. 9.6(a), and the following year G. B. Airy proposed the structure represented by Fig. 9.6(b). He visualized the crustal masses as logs, all of the same density, floating in water.

A log appearing higher out of the water than its neighbours must extend correspondingly deeper. Isostatic balance of the continents would be achieved by Pratt's principle if $d = t$ in Eq. (9.16), so that

$$\rho_c(h + t) = \rho_o t, \tag{9.17}$$

or by Airy's principle if $\rho_c = \rho_o$, in which case

$$\rho_c(h + d - t) = \rho_m(d - t). \tag{9.18}$$

Both principles contribute to the isostatic balance of continents and mountain ranges, the Airy principle being generally the more important.

Now we return to a consideration of Fig. 9.4. The continental crust is dominated by acid (Si rich) rocks that are less dense than the ocean floors, with which they are isostatically balanced. The continents do not spread out to cover the whole Earth uniformly, but are effectively swept clear of 60% of the surface area by the continuous renewal of the ocean floors. The continental material washed into the oceans as sediment is returned by underplating and via subduction zones and volcanism. What determines the thickness of the continental crust and thus the area that it occupies? As illustrated by Fig. 9.4, most of the continental area is close to sea level and this is a clue that it is reduced to that level by erosion, which has little effect on lowlands and continental margins. Higher elevations are maintained by tectonic activity, but erode rapidly, and so are relatively young. It is not fortuitous that most of the continental area is close to sea level, but a consequence of the erosion–regeneration cycle. Thus it is sea level itself that controls the continental thickness. This argument is used in Section 23.2 to support the assumption that, in the remote past, when less continental material had accumulated, its thickness was little different from the present time, but its area was smaller. The assumption is necessarily only an approximation, because, even assuming a constant volume of sea water, smaller continents would mean lower sea level.

9.4 Gravity anomalies and the inference of internal structure

Gravity anomalies are departures of observed gravity from the reference latitude variation in Eq. (6.38). They are of interest as indicators of internal structure, but it is important to recognise that there can be no unique solution to the problem of calculating internal density variations from observations of gravity on or above the surface. The forward problem of calculating gravity from any specified mass distribution is unambiguous, but there is, in principle, an infinite number of density distributions that could explain a gravity anomaly pattern. The range is restricted by plausibility, by setting up density models that mimic the observed gravity. This is aided by processing gravity data to represent anomalies in different ways. We refer to three types of anomaly presentation: free-air, Bouguer and geoid. The first two are mentioned in Section 6.3 and geoid anomalies are illustrated in Figs. 9.1 and 9.2. Here we take a closer look at the implications and assumptions, noting the essential point that gravity observations on a surface of variable elevation would be difficult to interpret directly and there are alternative ways of transposing them to the geoidal surface.

The free-air gradient means the radial variation of gravity above ground level. The global average value is $0.3086\,\mathrm{mGal\,m^{-1}}$ ($3.086 \times 10^{-6}\,\mathrm{s^{-2}}$). A free-air anomaly is the departure from the standard gravity formula (Eq. (6.38)) on the geoid (or mean sea level), calculated by assuming this gradient to apply from a surface point of observation. Usually, but not necessarily, the small latitudinal and local variations in the gradient are neglected. In principle it means calculating the gravity that would be observed on the geoid if all of the mass above it were collapsed below it. Of course, this is unsatisfactory in areas of more than very limited topography and topographic corrections are generally needed for the calculated anomalies to be useful. Then the resulting free-air anomalies are indications of concentrations or deficiencies in density below the points of observation. On scales smaller than 100 km or so the strength (or very high viscosity) of the lithosphere can support departures from isostatic balance that are apparent as free-air anomalies, notably at the margins of continents. This is an indication of what is known as the depth of compensation, that is the depth below which viscosity is low enough to equalize pressures (or below which homogeneity can be assumed). As we see from the global analysis in Section 9.3, on a larger scale isostasy prevails and this means equality of the masses in all vertical columns, so that free-air anomalies are weak. However, on a scale of several thousand kilometres free-air anomalies larger than those at intermediate scales are apparent from the geoid plot in Fig. 9.1. They are attributed to heterogeneity of the lower mantle. This is possible because, although the relatively low viscosity of the asthenosphere explains the isostatic balance at intermediate scales (200–2000 km) it does not nullify the effect of irregular masses in the more viscous lower mantle. These masses must be deep as they are not evident at the intermediate scale.

For the interpretation of local geological structures it is often more effective to calculate gravity on the geoid assuming complete removal of all material above it, instead of collapse to the geoid. This gives Bouguer anomalies. In the simplest cases, with no allowance for topography or heterogeneity, the removed material is assumed to be an extensive slab of uniform thickness equal to the height at the point of measurement. The gravity due to a slab of thickness h, density ρ and infinite horizontal extent, is (see Problem 9.2)

$$\delta g = 2\pi G\rho h, \tag{9.19}$$

and is independent of distance from it. Thus, calculation of gravity on the geoid by the Bouguer method means downward extrapolation from elevation h by a Bouguer gradient which is equal to the free-air gradient minus $\delta g/h = 2\pi G\rho$. Commonly a standard density, $2670\,\mathrm{kg\,m^{-3}}$, is used for this purpose and then the Bouguer gradient is $0.1967\,\mathrm{mGal\,m^{-1}}$ ($1.967 \times 10^{-6}\,\mathrm{s^{-2}}$), about 2/3 of the free-air gradient. If the density of the surface layers is known then it is obviously better to use that rather than the standard value and, as with free-air calculations, topographic corrections are usually necessary. Bouguer anomalies are of interest in studies of local crustal structure because they reflect density variations immediately below the geoid. On a continental scale Bouguer anomaly maps show systematic lows over continents because, by Eq. (9.19), they differ from free-air anomalies by $2\pi G\rho h$ at height h and on this scale isostatic balance prevails, making free-air anomalies small.

The purpose of free-air and Bouguer anomaly maps is to remove the effect of ground elevation, which obscures the underlying density variations. They are complementary, giving different information, and it is instructive to have both. The third method is to use geoid anomalies, that is variations in the height of the geoid, taking advantage of satellite observations of large-scale features, as in Fig. 9.1, and giving a different perspective on isostasy. The general idea can be understood from a simple example. If, in a broad, topographically featureless area of average density ρ, we have a patch with density $(\rho + \Delta\rho)$ underlying a layer of equal thickness but density $(\rho - \Delta\rho)$ then, by the argument in Section 9.3 the patch would be isostatically balanced. However, it would nevertheless appear as a geoid anomaly. The reason is that the deeper, denser layer gives enhanced gravity within and above it and, since this is the gradient of gravitational potential, in integrating upwards from the depth of compensation (below the patch) the potential corresponding to the geoid is reached at a lower level than in areas outside the patch. A density dipole of this form gives a geoid low. Conversely, if a denser layer overlies a less dense one it gives a geoid high. Thus a geoid anomaly gives information about the depth distribution of mass and not just its total. In this simple, plane layered model the geoid anomaly, ΔN (metres), is related to the depth dependence of the density anomaly by integrating Eq. (9.19) through it from the depth of compensation to the geoid potential,

$$\Delta N = -\frac{2\pi G}{g}\int_{-e}^{z_c} z\Delta\rho(z)\mathrm{d}z. \quad (9.20)$$

Equation (9.20) shows that geoid variation is due to the dipole moment of density above the depth of compensation, whereas gravity is caused by the integrated mass. For topography of a given elevation the depth of the mass contrast determines the associated geoid anomaly. Deeper masses give stronger geoid anomalies. Airy isostasy (Fig. 9.6b) is often associated with a crustal root of low-density material in the mantle, for example beneath the Himalaya and Andes the crust extends to depths of 75 km, giving geoid highs. In regions underlain by hot mantle, such as mid-ocean ridges and hot spot swells, as at Hawaii, and regions of continental extension with thin lithosphere, such as the Basin and Range of North America, the compensation occurs deeper and the geoid anomalies are stronger lows.

Measurements of the geoid by what we now know as astrogeodetic surveying have a history extending back into antiquity, being the original method of demonstrating that the Earth is approximately spherical. It determines the orientation of the vertical, that is the normal to the geoid, relative to the positions of stars. There was a rapid development of the method in the early 1700s, when the metre had been defined in terms of the dimensions of the Earth and parties of French scientists were sent to various parts of the world to measure it more precisely. At that time north–south surveys were much easier than east–west surveys because they are not critically dependent on precise timing of the east–west motions of stars, with the result that the ellipticity of the Earth was well determined. Gravity surveying developed from this work and especially from the ideas of M. Bouguer, who was surveying in South America in the 1730s and observed deflections of the vertical from a smooth geoid, caused by extreme topography in the Andes mountains. He attempted to measure the Newtonian gravitational constant, G, by two methods: one by observing the deflection of the vertical on the slopes of Mount Chimborazo and the other by comparing gravity on a plateau and on a low-lying plane, assuming the plateau to be a superimposed slab with a gravitational effect given by Eq. (9.19) (Bullen, 1975). He was thwarted by isostasy and realized that his results were unsatisfactory, but his work is recognized by identifying his name with the Bouguer method. It was an astrogeodetic survey across North India more than 100 years later that prompted the recognition of isostasy by J. H. Pratt, who led the survey party, and G. B. Airy, who, as British Astronomer Royal, had ultimate responsibility for the work in India.

The processes of interpreting geoid anomalies by Eq. (9.20) and calculating Bouguer anomalies use a flat-earth approximation and are satisfactory only on scales that are small compared with the radius of the Earth. The anomalies featured in Fig. 9.1 are too extensive for this approximation to be applied. It is notable that they are of greater amplitudes than more local anomalies, which are typically only a few metres, and this is an indication that they have a deep cause. Heterogeneity of the deep mantle is implicated and possibly even core–mantle boundary topography, although it is likely that the base of the mantle is soft enough for topography to be isostatically balanced with the heterogeneity in the D$''$ layer at the base of the mantle.

Particular interest in geoid anomalies arises from their relationship to dynamics of the mantle. Convection is driven by thermally generated density differences that must be reflected in the gravity field, but the problem is not a simple one. The mantle viscosity is strongly dependent on temperature, as well as being affected by pressure, and the convective pattern has no more than a remote resemblance to convection in an isoviscous fluid (Yoshida, 2004). There are also compositional heterogeneities to confuse the picture. Nevertheless, some conclusions are possible. Hager (1984) pointed out that zones of strong subduction are correlated with geoid highs, particularly when long-wavelength features of the geoid are emphasized by restricting consideration to low-degree harmonics, whereas the downwarping, obvious at the trenches, would be expected to cause lows. His explanation is that the topographic effect is masked by resistance to subduction in the deep mantle, causing a build-up of the dense subducting material responsible for the geoid highs and slowing the subduction, so that the lower-viscosity asthenosphere partially infills the topographic depressions. This requires viscosity to increase with depth in the mantle by a factor 30 or so, as most authors claim to be indicated also by post-glacial rebound (Sections 9.5 and 9.6).

9.5 Post-glacial isostatic adjustment

Delayed rebound of the Earth from depression by glacial ice has been recognized for many years, and Haskell (1935) showed that it can be applied to a calculation of mantle viscosity. He obtained a value of 10^{21} Pa s, which is still used as a satisfactory global average and is referred to as ‘the Haskell value’ (Mitrovica, 1996). More recent work, directed to the radial variation in viscosity, is reviewed by Peltier (1982, 2004), Lambeck (1990), Mitrovica (1996), Mitrovica and Forte (1997) and Kaufmann and Lambeck (2000). Information about the depth variation is conveyed by the geometrical form of the rebound and its lateral scale. We illustrate the principle with two simple models. One is essentially the Haskell model, an isoviscous half space, but with a superimposed lithosphere, and the other is a thin asthenosphere, within which all flow occurs, overlying a rigid mantle and with a very thin lithosphere that flexes freely but has no horizontal motion or viscous flow. The variations in the rate of rebound with lateral scale are quite different for these models. While they appear to be extremes, they are both deficient in not allowing for the flexural rigidity of the lithosphere. Although the lithosphere is defined as the cool and rigid surface layer of the mantle, this definition appeals to its temperature-dependent rheology. The effective thickness for post-glacial rebound is much less than the thickness apparent from the study of seismic waves, which stress the lithosphere on a time scale that is almost 10^{10} times shorter. We assess the lithospheric thickness in Section 20.4 and discuss its relevance to rebound at the end of this section.

Haskell’s (1935) estimate of mantle viscosity was based on a record of isostatic rebound of the glacially depressed area around the Gulf of Bothnia (Fennoscandia). The heaviest glaciation and the largest area of rebound are now recognized to be centred on Canada (Laurentia) and are illustrated by the vertical motions of North America, as plotted in Fig. 9.7. Peltier (2004) identified an area to the west of Hudson Bay as the location of the most rapid current rebound (1.5 cm/year) and inferred peak ice load (the Keewatin Dome), but GPS observations of rebound by Sella *et al.* (2007) put Hudson Bay itself at the centre of the action. Haskell (1935) considered the response of an isoviscous half space to a circular load with a Gaussian profile of rebound velocity, and matched his model to the Fennoscandian data, by a method referred to below. Availability of data from the larger-scale Laurentia rebound provides a measure of the scale dependence of rebound which is important to inferences about depth dependence of viscosity. Most recent analyses favour an asthenosphere with viscosity lower by at least an order of magnitude and a lower mantle with viscosity up to an order of magnitude higher than Haskell’s value. But there are strong lateral variations, so that no simple radial model can be unambiguous.

We now examine the simple, thin asthenosphere model of rebound. Both the upper and lower boundaries of the asthenosphere behave as rigid boundaries for the purpose of calculating the flow. We are modelling the present day flow and this represents a late stage in the isostatic adjustment to deglaciation, so we seek a self-consistent solution in which the local rebound speed, V_z, is everywhere proportional to the remaining isostatic anomaly in elevation, $-\zeta$. Departures from this state will have been subject to more rapid adjustment. Thus we can calculate a rebound time constant, τ, in terms of the scale of the flow pattern, where

$$V_z = -\zeta/\tau. \qquad (9.21)$$

The geometry of the flow is represented in Fig. 9.8. This shows the ‘moat’ of sinking land

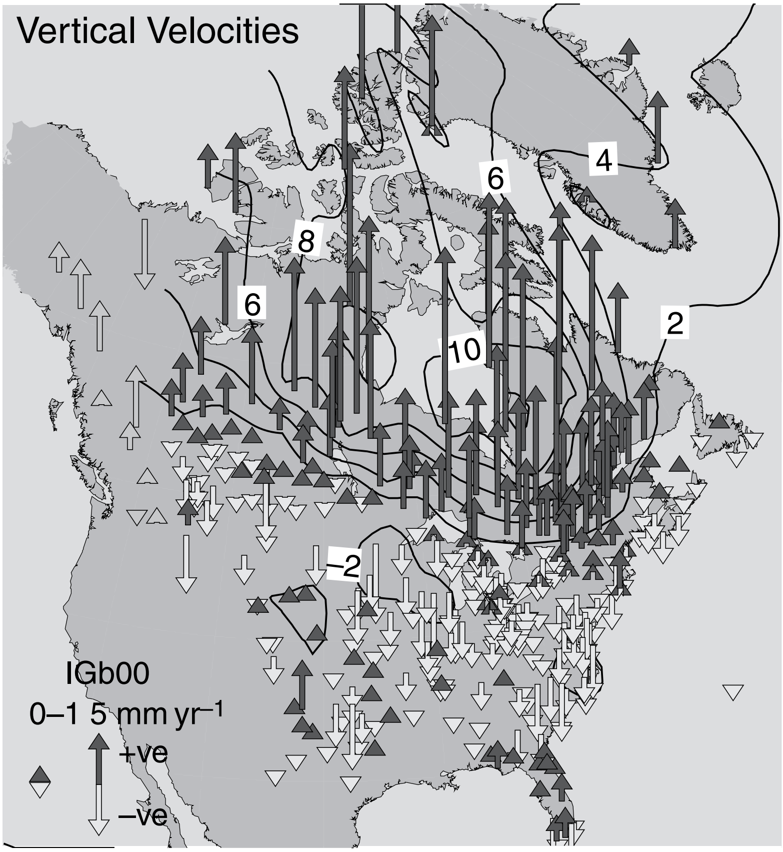


FIGURE 9.7 Laurentia glacial rebound from GPS measurements, showing a region of uplift centred near Hudson Bay, surrounded by a moat of subsidence at a boundary marked by a solid line. Reproduced, by permission, from Sella *et al.* (2007).

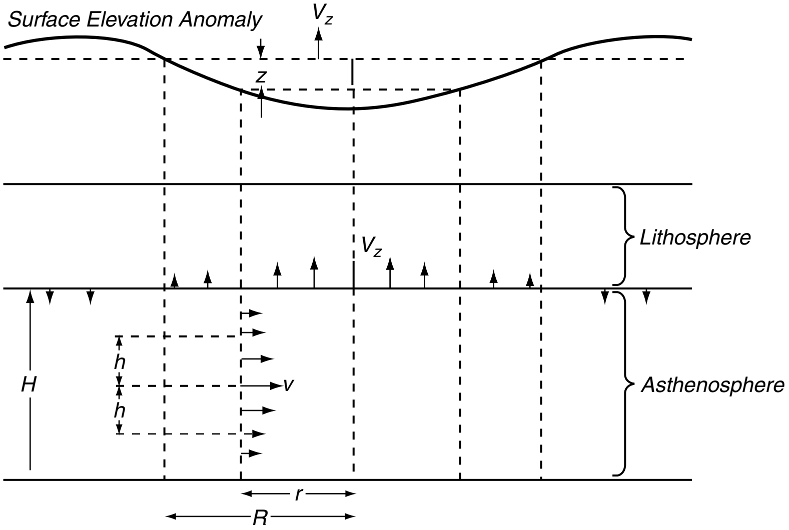


FIGURE 9.8 Model of rebound flow in a sharply bounded asthenosphere of uniform viscosity. The radius, *R*, of a circle of residual depression is much larger than the thickness, *H*, of the asthenospheric layer in which flow occurs. The lithosphere allows vertical movement by flexure, but does not permit horizontal flow.

surrounding the uplifting area, as required by the conservation of material. It is important to the approximations in this model that the horizontal scale is much larger than the vertical scale, so that the viscous drag of the horizontal flow limits the rebound rate. The local driving force for the flow is the deficiency in hydrostatic pressure,

$$P = \rho g \zeta, \tag{9.22}$$

where $\rho = 3400\,\mathrm{kg\,m^{-3}}$ is the asthenospheric density, because it is asthenospheric material that is displaced, and $g = 9.9\,\mathrm{m\,s^{-2}}$ is the value of gravity at that level. The model has circular symmetry, so that the asthenospheric flow is radially inwards.

The flow through a central annulus of radius r in the asthenosphere is driven by the pressure gradient, $\mathrm{d}P/\mathrm{d}r$. Then the shear stress, $\sigma(h)$, on each of the upper and lower faces of a central layer of thickness $2h$, as in Fig. 9.8, is due to this pressure gradient acting over the thickness $2h$,

$$\sigma(h) = \frac{(\mathrm{d}P/\mathrm{d}r)2h2\pi r}{2.2\pi r} = \frac{\mathrm{d}P}{\mathrm{d}r}h. \tag{9.23}$$

Therefore the flow speed, v, at (h, r) is given by

$$\frac{\mathrm{d}v}{\mathrm{d}h} = -\frac{\sigma(h)}{\eta} = \frac{\mathrm{d}P}{\mathrm{d}r}\frac{h}{\eta} = -\rho g\frac{\mathrm{d}\zeta}{\mathrm{d}r}\frac{h}{\eta}, \tag{9.24}$$

where η is the viscosity and Eq. (9.22) gives the substitution for P. Integrating to level h from one boundary of the asthenosphere at $H/2$,

$$v = \frac{\rho g}{\eta}\frac{\mathrm{d}\zeta}{\mathrm{d}r}\left(\frac{H^2}{8} - \frac{h^2}{2}\right), \tag{9.25}$$

and hence the total volume flow of material though the annular ring of radius r is

$$F = 2\pi r \cdot 2\int_0^{H/2} v\,\mathrm{d}h = \frac{\pi}{6}H^3\rho g r\frac{\mathrm{d}\zeta}{\mathrm{d}r}. \tag{9.26}$$

The rate of rise of material at radius r is

$$V_z = \frac{1}{2\pi r}\frac{\mathrm{d}F}{\mathrm{d}r} = \frac{H^3\rho g}{2\eta r}\left(\frac{\mathrm{d}\zeta}{\mathrm{d}r} + r\frac{\mathrm{d}^2\zeta}{\mathrm{d}r^2}\right), \tag{9.27}$$

so that, by assuming that the form of the residual depression is preserved, as its amplitude decreases, and therefore that we can substitute for V_z by Eq. (9.21), we have the differential equation for $\zeta(r)$,

$$\frac{\mathrm{d}^2\zeta}{\mathrm{d}r^2} + \frac{1}{r}\frac{\mathrm{d}\zeta}{\mathrm{d}r} + \left[\frac{12\eta}{\rho g H^3 \tau}\right]\zeta = 0. \tag{9.28}$$

This is the special case (for $n = 0$) of Bessel's equation

$$\frac{\mathrm{d}^2\zeta}{\mathrm{d}x^2} + \frac{1}{x}\frac{\mathrm{d}\zeta}{\mathrm{d}x} + \left[1 - \frac{n^2}{x^2}\right]\zeta = 0, \tag{9.29}$$

where

$$x = \left[\frac{12\eta}{\rho g H^3 \tau}\right]^{1/2} r. \tag{9.30}$$

The solution is therefore the zero-order Bessel function

$$\zeta = \zeta(r = 0)\,J_0(x). \tag{9.31}$$

The first zero of this function occurs at $x = 2.4$, so that the radius of the circular depressed area, R, is given by

$$\left[\frac{12\eta}{\rho g H^3 \tau}\right]^{1/2} R = 2.4. \tag{9.32}$$

For this model, the time constant for the approach to isostatic equilibrium is therefore

$$\tau = 2.1\eta R^2/\rho g H^3 \tag{9.33}$$

and, if we make the unknown η the subject of the equation,

$$\eta = 0.48\rho g H^3 \tau / R^2. \tag{9.34}$$

By this model, the estimated asthenospheric viscosity depends very strongly on its assumed thickness, but in reality it is not sharply bounded.

The other extreme model is an isoviscous half space, similar to that of Haskell (1935) but with the addition of a lithosphere that acts as a frictional boundary and moves only vertically, not participating in the horizontal flow. Haskell considered a depression of Gaussian form and, by Fourier–Bessel analysis, resolved it into a sum (more correctly, integral) of Bessel functions, each of which retains its form as it decays and so can be identified with a relaxation time. The decay of the whole depression is then obtained from the sum of the individually decaying Bessel functions, J_0, so that it progressively approaches the form of the most slowly decaying term. Thus the solution is superficially similar to that of the bounded asthenosphere model above, with the most slowly decaying Bessel term dominating the geometry at an advanced stage in the recovery. But there is a crucial difference that we can see by a simple dimensional analysis. As in Eq. (9.33), we require a relaxation time, τ, proportional to $(\eta/\rho g)$. H does not enter the problem, so the only other dimensioned quantity involved is the radius, R, of the depression. Thus, to balance dimensions we must have

$$\tau = C\eta/\rho g R, \tag{9.35}$$

where C is a dimensionless constant. The two extreme models lead to very different dependences of τ on R, proportionality to R^2 for the bounded asthenosphere and $1/R$ for the uniform half space.

The areas of Fennoscandian ($R \approx 600$ km) and Laurentide ($R \approx 1300$ km) rebound are sufficiently different that we can use them as a test of Eqs. (9.34) and (9.35). Since the equations represent extreme models, and we expect the truth to lie somewhere between them, we write $\tau \propto R^k$, with $k = 2$ for the thin asthenosphere and $k = -1$ for the half space. For Fennoscandia there is good agreement on the relaxation time, 4600 years (Mitrovica, 1996) or 4350 years (Peltier, 1998), but for Laurentia the estimates are seriously divergent: 6700 years (Mitrovica) or 3400 years (Peltier). The Mitrovica numbers give $k = 0.49$, midway between the extremes, indicating that, although the asthenosphere is not sharply bounded and the deep mantle participates in the rebound, there is a strong increase in viscosity with depth. The Peltier values give $k = -0.32$ and are, therefore, indicative of a much more nearly isoviscous mantle. Although the models are greatly simplified, these conclusions coincide with more detailed analyses. Mitrovica and Forte (1997) and Kaufman and Lambeck (2000) concluded that the increase in viscosity between the uppermost and lower mantles is of order a factor 100, whereas Peltier (1998) concluded that the increase at 660 km depth is no greater than a factor of 5. We need to consider the reason for such different conclusions, drawn from what are essentially the same observations.

As we have mentioned, the simple models above do not account satisfactorily for the lithosphere, which imposes a constraint on the motion in the underlying mantle that depends on the scale of the motion. This scale dependence is determined by the flexural rigidity of the lithosphere and is a strong function of its thickness. We discuss the lithospheric thickness in Section 20.4 and, for flexure on a time scale of a few thousand years in the continental areas where rebound is observed, conclude that it is about 60 km. However, Peltier favoured 120 km and Lambeck *et al.* (1996) pointed out that the thick lithosphere assumed by Peltier accounts for the relatively slight viscosity contrast across the mantle in his models and a high estimate of upper mantle viscosity.

The scale dependence problem draws attention to the manner in which rebound calculations are carried out. The current upward surface motion is identified not with a residual surface depression but with the history of the ice load that caused it. The viscosity models rely on ice history models and are not derived from instantaneous measures of relaxation time constants. If we consider the different spatial harmonic terms of a rebound pattern (the Fourier–Bessel coefficients in the Haskell calculation), then they have different relaxation times, so that the present pattern depends not only on the history of the ice load, but on the response to it. Traditionally, rebound has been observed as the rise of land relative to sea level, by identification of former shore lines, now elevated. With allowance for the rise in sea level itself, this records the history of the process at specific localities. Now the Grace satellite presents the opportunity to observe the present rate of rise globally, and not just at shore lines. This will provide a resolution of the question of the adequacy of the observed rebound to explain the variation in ellipticity, considered in the following section.

Viscosity is a controlling parameter in mantle convection (Chapter 22). In this case the lateral variations are more significant because subduction of the cool, almost rigid lithospheric slabs requires flow in the adjacent mantle. As with rebound, the flow occurs most readily in layers of lowest viscosity and a detailed three-dimensional flow pattern is difficult to model on a sufficiently fine scale. However, information about viscosity derived from convection studies is complementary to that derived from rebound. The mechanical energy dissipation is thermodynamically related to the heat flux and so gives a well constrained value of $\int \dot{\varepsilon}^2 \eta \mathrm{d}V$ through the volume of the mantle, where $\dot{\varepsilon}$ is the strain rate. The model of an almost rigid lithosphere overlying a weak asthenosphere with strongly increasing viscosity a greater depths is consistent with the discussion of convective energy in Section 13.2.

9.6 Rebound and the variation in ellipticity

Ice-age glaciation was a high latitude phenomenon, so that rebound means a slow withdrawal of mantle material from lower latitudes to infill the high latitude depressions. A result is a slow decrease in the gravitational potential coefficient J_2 (see Eqs. (6.1), (6.2) and (6.14)) and, as discussed in Section 6.4, there is a corresponding effect on the rotation rate, ω. First, we consider the consequence of supposing that the excess J_2 is a global phenomenon, that is, the Earth as a whole is flattened. Then J_2 is related to the surface ellipticity, $f=(a-c)/a$, by Eq. (6.20), where a, c are equatorial and polar radii, and for the present purpose we can ignore m in this equation by assuming $\dot{J}_2 = (2/3)\dot{f}$. If the surface deformation is ellipsoidal we can write it as

$$\Delta r = -\Delta J_2 a(3\cos^2\theta - 1)/2, \tag{9.36}$$

where ΔJ_2 is the excess J_2 attributable to polar depression. Then the gravitational energy of the deformation is an integral over the surface area A,

$$E = 1/2 \int g\rho(\Delta \mathrm{r})^2 \mathrm{d}A = (2\pi/5)a^4\rho g(\Delta J_2)^2, \tag{9.37}$$

and the energy dissipation by the decreasing J_2 is

$$\mathrm{d}E/\mathrm{d}t = (4\pi/5)a^4\rho g(\Delta J_2)\dot{J}_2. \tag{9.38}$$

The strain rate is simply $\dot{f}$ and so we can write it as

$$\dot{\varepsilon} = \dot{f} = (3/2)\dot{J}_2, \tag{9.39}$$

and then the dissipation is, in terms of the mean or effective viscosity, $\bar{\eta}$,

$$\begin{aligned}\mathrm{d}E/\mathrm{d}t = \int \dot{\varepsilon}^2\eta \mathrm{d}V &\approx (9/4)(\dot{J}_2)^2\bar{\eta}(4\pi/3)(r^3_{\mathrm{mantle}} - r^3_{\mathrm{core}}) \\ &= 7.9a^3(\dot{J}_2)^2\bar{\eta}\ .\end{aligned} \tag{9.40}$$

Equating (9.38) and (9.40),

$$\eta = 0.32\rho g\ a[\Delta J_2/\dot{J}_2] = 2.2\times 10^{18}\tau(\text{years}) \tag{9.41}$$

for η in Pa s, because $[\Delta J_2/\dot{J}_2]$ is the relaxation time, τ. If we assume $\eta = 10^{21}$ Pa s, the Haskell value, then $\tau \approx 450$ years. This just confirms that, for the isoviscous model, the relaxation time decreases with scale and that, by this model, a depression of global extent would now have recovered to the point of being unobservable. On the other hand if we apply the simplistic compromise model, $\tau \propto R^k$ with $k = 0.5$, then the global scale relaxation time would be about 13 000 years, and if there were ever a significant global scale depression then it would now be dominant. Neither of these alternatives appears plausible, so we consider the consequence of attributing the J_2 variation to the sum of the effects of more local rebound.

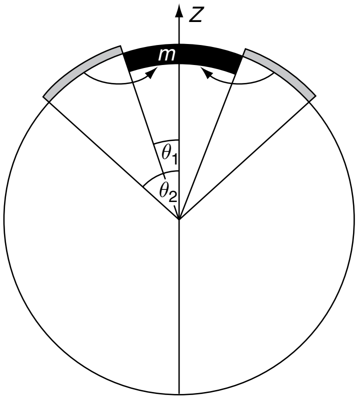


FIGURE 9.9 The rebound if Laurentia is modelled as the development of a uniform spherical cap of angular radius θ_1 and mass m by withdrawal of this mass from the annular ring extending from θ_1 to θ_2.

Since Laurentia is the largest of the well-studied areas of glaciation and rebound, we use it to estimate the ellipticity change that is explained by it. The effect to be explained is a long-term average rate of change $\mathrm{d}J_2/\mathrm{d}t \equiv \dot{J}_2 = -2.8 \times 10^{-11}\,\mathrm{year}^{-1}$ and, if Laurentia accounted for a large part of this, we could be confident that the process is understood. Its rebound is modelled, in the manner of Fig. 9.9, as the development of a uniform spherical cap of angular radius θ_1 and mass m by withdrawal of this mass from an annular ring extending from θ_1 to θ_2. The moment of inertia of the cap about its axis (z) is

$$\Delta \dot{I}_{z1} = ma^2(2 - \cos\theta_1 - \cos^2\theta_1)/3 \tag{9.42}$$

and the moment of inertia of the annular ring about the same axis is

$$\Delta I_{z2} = ma^2(3 - \cos^2\theta_1 - \cos\theta_1\cos\theta_2 - \cos^2\theta_2)/3. \tag{9.43}$$

Thus the total change in moment of inertia about the z axis due to the shift of the mass m is

$$\begin{aligned}\Delta I_z = \Delta I_{z1} - \Delta I_{z2} = -ma^2(1 + \cos\theta_1 \\ - \cos\theta_1\cos\theta_2 - \cos^2\theta_2)/3.\end{aligned} \tag{9.44}$$

We calculate the changes in moment of inertia about the x and y axes by applying the rule that for any distribution of mass m on a spherical surface the sum of the moments of inertia about three mutually perpendicular axes is

$$I_x + I_y + I_z = 2mR^2. \tag{9.45}$$

(This is seen by writing the moments of inertia of any mass element Δm at x, y, z about the three axes as $\Delta I_x = \Delta m(y^2 + z^2)$, $\Delta I_y = \Delta m(x^2 + z^2)$, $\Delta I_z = \Delta m(x^2 + y^2)$, so that $\Delta I_x + \Delta I_y + \Delta I_z = 2\Delta m(x^2 + y^2 + z^2) = 2\Delta mR^2$, noting that this applies to all mass elements individually and therefore to any distribution of mass on a spherical surface.) In the rebound situation we are considering only a redistribution of mass with no change in its total, so that $(\Delta I_x + \Delta I_y + \Delta I_z) = 0$ and, with symmetry about z,

$$\Delta I_x = \Delta I_y = -\Delta I_z/2 \tag{9.46}$$

with ΔI_z given by Eq. (9.44).

We can allow for a misalignment of z with the rotational axis, c. Let the z axis be at co-latitude θ in the (c – a) plane. Then

$$\begin{aligned}\Delta I_c &= \Delta I_z\cos^2\theta + \Delta I_x\sin^2\theta \\ &= \Delta I_z(\cos^2\theta - (1/2)\sin^2\theta),\end{aligned} \tag{9.47}$$

$$\begin{aligned}\Delta I_a &= \Delta I_z\sin^2\theta + \Delta I_x\cos^2\theta \\ &= \Delta I_z(\sin^2\theta - (1/2)\cos^2\theta),\end{aligned} \tag{9.48}$$

$$\Delta I_b = \Delta I_y = -(1/2)\Delta I_z. \tag{9.49}$$

Here a, b are equatorial axes that are averaged for the purpose of calculating J_2, so that

$$\begin{aligned}\Delta J_2 &= [\Delta I_c - (1/2)(\Delta I_a + \Delta I_b)]/Ma^2 \\ &= \Delta I_z(9\cos^2\theta - 3)/4Ma^2,\end{aligned} \tag{9.50}$$

where M is the total Earth mass and ΔI_z is given by Eq. (9.44). Thus, for a rate of mass transfer $\dot{m}$ into a cap at latitude θ,

$$\begin{aligned}\dot{J}_2 = -(\dot{m}/M)(1 + \cos\theta_1 - \cos\theta_1\cos\theta_2 \\ - \cos^2\theta_2)(3\cos^2\theta - 1)/4.\end{aligned} \tag{9.51}$$

Using data by Peltier (2004), we estimate $\dot{m} = 2.1 \times 10^{14}\,\mathrm{kg/year}$ (of material of asthenospheric density $3400\,\mathrm{kg\,m^{-3}}$) into a Laurentide cap of angular radius $\theta_1 = 15^\circ$ at co-latitude $\theta = 30^\circ$. The contribution to $\dot{J}_2$ depends on what is assumed for θ_2. A thin asthenosphere with rapidly increasing viscosity at greater depth would restrict θ_2 to a band of order $2\theta_1 = 30^\circ$. If this value is assumed then the contribution to $\dot{J}_2$ is -4.2×10^{-12} year^{-1}, 15% of the total. An annular ring twice as wide, that is $(\theta_2 - \theta_1) = 30^\circ$, uld give $\dot{J}_2, = -8.6 \times 10^{-12}$ year^{-1}, still only 31% of the total. If Laurentia is correctly modelled, then it is not as dominant as supposed. Adding Fennoscandia would make little difference, because not only is the mass flow smaller, but so is its angular spread. Antarctic rebound is not as well documented and may be more significant than has been recognized, particularly if the major ice retreat there occurred much later than in the northern hemisphere, so that rebound is at an earlier, faster stage. The discrepancy remains to be resolved.

10

Elastic and inelastic properties

10.1 Preamble

We refer to a material as elastic if it may be deformed by stress but recovers its original size and shape when released from the stress. In fact no material is perfectly elastic, even if subjected to indefinitely small stresses, but if recovery is almost complete (say $>99\%$) and, subject only to inertial delay, effectively instantaneous, a material is regarded as elastic. Even for such a material we recognize that the slight departure from perfect elasticity has important consequences, which include the attenuation of seismic waves. At high shear stresses departure from ideal elasticity increases sharply. Each material has an approximate elastic limit or yield point, which is the magnitude of stress above which inelastic or permanent deformation starts to become significant. This is additional to the recoverable elastic response and may increase with time at constant stress. There is no sharp cut-off so that very prolonged stresses can cause continued very slow deformation or creep, especially at high temperatures, as in mantle convection.

Small elastic strains are normally proportional to stress (Hooke's law). Then the ratio of stress to strain is an elastic ('stiffness') modulus. Stress is force per unit area, measured in pascals ($\mathrm{Pa} \equiv \mathrm{Nm}^{-2}$), and strain is a fractional change in some dimension or dimensions, so that elastic moduli have the same units as stress, i.e. pascals. The theory of elasticity, as we normally consider it, deals only with very small strains, for which elastic moduli are effectively material constants. This is infinitesimal strain theory. It describes elastic deformations in response to stresses that are very small compared with values of the elastic moduli. Pressure in the deep interior of the Earth is much too large for this assumption to be even approximately valid. In this situation an elastic modulus must be defined as the ratio of a small change in stress to the consequent small strain increment, superimposed on a much larger ambient compression. The moduli all increase systematically with pressure and a more general, finite strain theory is required. This is a subject of Chapter 18.

Elasticity theory is simplest for materials that are isotropic, that is having properties that are the same in all directions. They include polycrystalline materials with randomly oriented crystals, such as undeformed metals or igneous rocks, as well as glassy or amorphous substances. Elastically isotropic materials have two independent elastic moduli. There are several ways of choosing these moduli, depending on the form of the strain to be represented, but since only two can be independent it follows that any one of them can be represented in terms of any other two, as summarized in Appendix D. In seismology and much other geophysics we generally prefer to use bulk modulus, K, and rigidity modulus, μ, as the two independent moduli. The particular reason for geophysical interest in bulk modulus is that it provides a link between seismologically observed elastic wave speeds and the density profile of the Earth. With a minor proviso concerning the elasticity of a composite (Section 10.4), the combination of P-wave speed, controlled by the modulus χ (Eq. (10.5)), and shear wave speed, controlled by rigidity, μ, provides a direct observation of the ratio $K/\rho = \mathrm{d}P/\mathrm{d}\rho$, where ρ is density and P is pressure.

A pure shear strain with no change in volume, described by μ, involves no change in temperature and μ is unambiguous. On the other hand, a pure compression with no change in shape, which is described by K, requires specification of the conditions under which the compression is applied. If temperature is maintained constant we have the isothermal modulus, K_T, but if the material is thermally isolated or if the compression is too rapid to allow any transfer of heat, as in a seismic P wave, then the temperature rises and a higher, adiabatic modulus, K_S, applies. This allows for the fact that compression is partially offset by thermal expansion. The two principal bulk moduli, K_T and K_S, are related by an identity

$$K_S = K_T(1 + \gamma\alpha T) \qquad (10.1)$$

(T is temperature, α is volume expansion coefficient and γ is the dimensionless Grüneisen parameter, which has numerical values between 1.0 and 1.5 in the Earth). In the deep Earth the difference between K_S and K_T is between 4% and 10%. The other moduli, considered in the following section, all describe deformations involving changes in both volume and shape, so that relationships between isothermal and adiabatic versions are inconveniently complicated. Equation (10.1) is one of the thermodynamic identities that link bulk modulus to other properties, such as thermal expansion (Appendix E). Rigidity (and therefore also the other moduli that can be regarded as combinations of K and μ, as in Appendix D) are not subject to comparable controls. This means that, when we consider effects such as the temperature dependences of moduli (Section 17.5), rigidity cannot be treated with the same thermodynamic rigour as bulk modulus and further assumptions are needed.

For many purposes isotropy is an adequate approximation, but elastic anisotropy is well recognized in both the mantle and the solid inner core. It can arise both from an alignment of crystals, which are individually anisotropic, and from the layering of materials with different properties. The simplest crystals are those with cubic structures, which have three independent moduli, but more general anisotropies, requiring a larger number of moduli, are normal in minerals. The most general situation of a crystal completely lacking structural symmetry is represented by 21 moduli (for a reason outlined in Section 10.3). Although mineralogists face such complications, in the Earth the form of anisotropy most commonly considered in detail is uniaxial, that is, having properties in one direction differing from those in the perpendicular plane, within which there is no variation. In the upper mantle the single axis is assumed to be vertical (radial) and seismologists refer to such a structure as transversely isotropic. It is represented by five moduli (Section 10.3). Although azimuthal anisotropy occurs, both on continents (e.g. Davis, 2003) and on the ocean floors, on a global scale gravity selects the vertical axis as unique. Transverse anisotropy is smeared out statistically, so that transverse isotropy is to be expected as a global average. In the inner core it appears that the rotational axis is selected as a unique direction. This is best explained by the precipitation of added material predominantly on its equator and deformation towards the equilibrium ellipticity (Yoshida *et al.*, 1996).

Our understanding of the Earth's interior is derived almost entirely from seismology and its interpretation, based on elastic and inelastic properties. These properties include anisotropy, the behaviour of composites and polycrystalline materials, effects of temperature and pressure and the frequency variation of elasticity, all of which need to be understood to make maximum use of seismological data.

10.2 Elastic moduli of an isotropic solid

The simplest derivation of the relationships between moduli starts with Young's modulus, E, and Poisson's ratio, ν, which are defined with the other moduli in Table 10.1. Consider a rectangular prism, as in Fig. 10.1, with axes x, y, z, chosen to be parallel to the principal stresses, σ_x,

Table 10.1 Definitions of elastic moduli for an isotropic solid, referred to the cube in Fig. 10.1. Directions of axes 1, 2, 3 are directions of principal stresses, σ, as indicated for the case of shear modulus in the figure. Strains, ε, are fractional changes in dimension. K, μ, E, λ and χ are moduli relating stress to strain, with dimensions force per unit area, pascals (Pa $\equiv$ Nm^{-2}). Poisson's ratio, ν, is a dimensionless ratio of strains (lateral/axial) for a body stressed only in the axial direction

Modulus	Symbol (alternatives)	Definition(s)
Bulk modulus (incompressibility)	$K\,(B)$	$-\dfrac{V\Delta P}{\Delta V}\,(V=\text{volume},\ P=\text{pressure})$ $=\dfrac{\sigma_1}{3\varepsilon_1}$ with $\sigma_2=\sigma_3=\sigma_1=P$ $\varepsilon_2=\varepsilon_3=\varepsilon_1=\dfrac{1}{3}\Delta V/V$
Rigidity (shear modulus)	$\mu\,(G)$	$\dfrac{\sigma_s}{\varepsilon_s}$ (as in Fig.10.1) $=\dfrac{\sigma_1}{2\varepsilon_1}$ for $\sigma_s=\sigma_1=-\sigma_2,\ \sigma_3=0$ $\varepsilon_s=2\varepsilon_1=-2\varepsilon_2,\ \varepsilon_3=0$
Young's modulus	$E\,(q,Y)$	$\dfrac{\sigma_1}{\varepsilon_1}$ for $\sigma_2=\sigma_3=0$
Poisson's ratio	$\nu\,(\sigma)$	$\dfrac{-\varepsilon_2}{\varepsilon_1}=\dfrac{-\varepsilon_3}{\varepsilon_1}$ for $\sigma_2=\sigma_3=0$
Lamé parameter	λ	$\dfrac{\sigma_1}{(\varepsilon_2+\varepsilon_3)}$ for $\varepsilon_1=0$
Modulus of simple longitudinal strain (axial modulus)	$\chi\,(m)$	$\dfrac{\sigma_1}{\varepsilon_1}$ for $\varepsilon_2=\varepsilon_3=0$

σ_y, σ_z, that is σ_x is the force per unit area exerted on a face normal to x and there are no shear stresses on the faces of this prism. The strains, ε_x, ε_y, ε_z, are the fractional changes in dimension in the three directions. First impose the stress σ_x only with no transverse stresses ($\sigma_y=\sigma_z=0$). Then

$$\varepsilon_x=\sigma_x/E, \tag{10.2}$$

where E is Young's modulus, and

$$\varepsilon_y=\varepsilon_z=-\nu\varepsilon_x=-\nu\sigma_x/E, \tag{10.3}$$

with ν being Poisson's ratio. The responses to additional stresses are simply additive, so that for three arbitrary principal stresses

$$\begin{aligned}\varepsilon_x&=\left[\sigma_x-\nu\left(\sigma_y+\sigma_z\right)\right]/E,\\ \varepsilon_y&=\left[\sigma_y-\nu\left(\sigma_x+\sigma_z\right)\right]/E,\\ \varepsilon_z&=\left[\sigma_z-\nu\left(\sigma_x+\sigma_y\right)\right]/E.\end{aligned} \tag{10.4}$$

By imposing particular conditions on Eqs. (10.4) we can relate the other moduli to E and ν. If $\sigma_1=\sigma_2=\sigma_3$ the stress is a hydrostatic pressure increment, $-\Delta P$, and the corresponding fractional volume change is $\Delta V/V=(1+\varepsilon)^3-1\approx 3\varepsilon$, as in Table 10.1. Treatment of the rigidity modulus is illustrated in Fig. 10.1. The compressions and rarefactions in a compressional wave depend on the lateral extent of the body in which it is propagating. In a rod that is thin compared with the wavelength, the extensions and

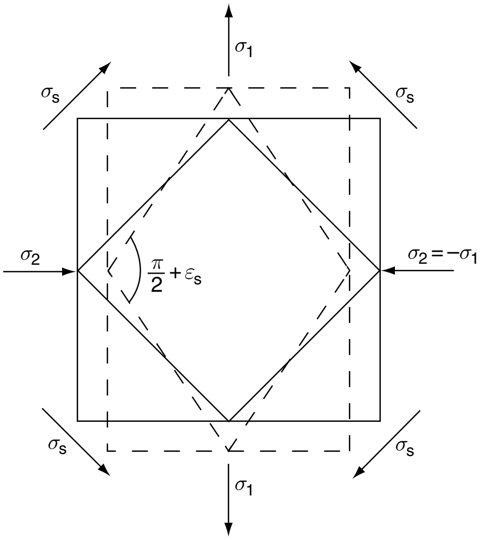


FIGURE 10.1 Shear deformation, referred to the axes of the principal stresses, σ_1 and σ_2. Broken lines represent strained cubes. Shear strain, $\varepsilon_s = \varepsilon_1 - \varepsilon_2 = 2\varepsilon_1$, where $\varepsilon_2 = -\varepsilon_1$ are the linear strains in the directions of the principal stresses. The shear stress, σ_s, is represented by tangential forces acting on the faces of the inner cube and is equal in magnitude to σ_1.

compressions are described by Young's modulus, because there is no lateral stress. However, in the Earth the wavelengths of seismic waves are normally very short compared with the extent of the medium in which they propagate. In this situation the juxtaposition of alternate half wavelengths of compression and extension prevents the lateral strains that occur as a Poisson's ratio effect in a thin rod. Then the deformations in the direction of wave propagation are simple longitudinal strains, with no lateral strain. The relevant modulus, χ, is obtained by putting $\chi = \sigma_x/\varepsilon_x$ with $\varepsilon_y = \varepsilon_z = 0$ in Eqs. (10.4). The Lamé parameter, λ, which is used in the tensor representation of stress and strain in Chapter 11, may be defined as the ratio of normal stress in a direction of zero strain to the areal strain of a perpendicular plane. By relating each of the other moduli to E and ν, we can write any of the moduli in terms of any other two, as in Appendix D. The most important of these relationships for seismology is

$$\chi = K + (4/3)\mu. \tag{10.5}$$

This is the modulus that controls the speed of compressional waves in the Earth.

10.3 Crystals and elastic anisotropy

As mentioned in Section 10.1, an isotropic material is one in which the component crystals are randomly oriented. The elasticities of individual crystals cannot be described by just two moduli, such as K and μ, but require three (for cubic crystals) or more, depending on crystal symmetry. Anisotropy of a polycrystal is observed if the alignment of component crystals is non-random; this is normally observed in deformed materials. It is seen in the Earth's upper mantle and in the inner core. Anisotropy may also be caused by layering, even if the individual layers are composed of isotropic materials (Problem 10.2, Appendix J).

To represent anisotropic properties we need a system for identifying components of strain with different orientations and this is achieved by treating them as spatial variations in the displacements of material points. If a crystal is stretched in the x direction then the x displacement increases with x and the corresponding strain component is represented by ε_{xx}. If the x displacements of material points vary with y then we have a strain component ε_{yx} and so on. But shear components of strain, such as ε_{yx}, are changes in angle so that $\varepsilon_{yx} = \varepsilon_{xy}$ and we are really interested in the total change in angle, that is $(\varepsilon_{yx} + \varepsilon_{xy})$. Therefore, there are just six independent strain components. They are represented by the symbol e to distinguish them from ε, that is $e_{xx} \equiv \varepsilon_{xx}$, $e_{yy} \equiv \varepsilon_{yy}$, $e_{zz} \equiv \varepsilon_{zz}$, $e_{xy} \equiv \varepsilon_{xy} + \varepsilon_{yx}$, $e_{yz} \equiv \varepsilon_{yz} + \varepsilon_{zy}$, $e_{xz} \equiv \varepsilon_{xz} + \varepsilon_{zx}$. The first three of these are linear strains and the last three are shear strains. This notation is general and is used also in the rock mechanics discussion in Chapter 11. It can be applied to either isotropic or anisotropic materials. The difference is in the relationships between these strain components and the correspondingly identified stress components. In the case of anisotropy there are cross terms, such as the appearance of a strain e_{xy} in response to a stress σ_{xx}, that do not occur with isotropic materials.

This is the basis for the notation used to identify the elastic moduli of crystals. There are six components of stress, each of which may cause

Table 10.2a Olivine elastic constants c_{ij} (GPa) with orthorhombic symmetry (nine constants). 300 K values for $(Mg_{0.9}Fe_{0.1})_2SiO_4$ from Anderson and Isaak (1995)

320.6	69.8	71.2			
69.8	197.1	74.8			
71.2	74.8	234.2			
			63.7		
				77.6	
					78.3

Table 10.2b Average elastic constants (GPa) for the upper 196 km of the PREM model of the mantle, which is transversely isotropic (elastically equivalent to hexagonal symmetry). There are five independent constants because isotropy in the horizontal plane gives $c_{66} = (1/2)(c_{11}-c_{12})$

224	84.8	85.4			
84.8	224	85.4			
85.4	85.4	212			
			65.7		
				65.7	
					69.6

six components of strain and the six simultaneous equations that relate them therefore require 36 coefficients (elastic moduli, c). They are given numerical suffixes according to the convention

$$1 \equiv xx,\ 2 \equiv yy,\ 3 \equiv zz,\ 4 \equiv yz,\ 5 \equiv zx,\ 6 \equiv xy,$$

bearing in mind that $xy = yx$ etc. Hence stress and strain are related by six equations of the type

$$\sigma_{xx} = c_{11}e_{xx} + c_{12}e_{yy} + c_{13}e_{zz} + c_{14}e_{yz} + c_{15}e_{zx} + c_{16}e_{xy}. \tag{10.6}$$

There is some redundancy in the coefficients because, as for $e_{xy} = e_{yx}$ above, $c_{ij} = c_{ji}$, so that the 30 coefficients for which $i \neq j$ are reduced to 15 independent moduli, making 21 altogether. The problem of measuring elasticity as general as this is forbidding and, for minerals of interest, crystal symmetries reduce the number of independent moduli. Values for a large number of minerals at laboratory pressure and temperature are catalogued by Bass (1995) according to their structures: monoclinic (13 moduli), orthorhombic (nine moduli), trigonal and tetragonal (six or seven moduli), hexagonal (five moduli) and cubic (three moduli). Temperature variations are reported by Anderson and Isaak (1995).

Olivine crystals are orthorhombic (Table 10.2a) and their alignment by shear in the upper mantle gives rise to anisotropy of seismic wave speeds. The reference Earth model PREM treats the uppermost 196 km of the mantle as transversely isotropic, meaning that it is azimuthally isotropic, with the same properties in all horizontal directions, but with different properties vertically. This requires five independent elastic constants (averages are given in Table 10.2b). This is probably a good assumption for the global average but inadequate for regional studies where a lower symmetry is required, e.g. orthorhombic in regions where azimuthal anisotropy has been measured (Davis, 2003).

For an idea of the effect of crystal symmetry and relationships to the isotropic moduli in Section 10.2, consider the simplest case, cubic crystals. Strain energy is proportional to the product of stress and strain (and therefore to the square of strain) for small strains, as considered in this chapter, and may be calculated as $(1/2)\sigma e$ using the series of equations (10.6) etc. But we retain only terms that remain the same with all possible rotations or interchanges of the equivalent cubic axes, and there are four body diagonals of a cube about which rotations by $2\pi/3$ interchange x, y and z. Many of the terms in the product (σe) appear with reversed signs with one or more such rotations and are therefore discounted. Others appear as identical in the strain components but with different c_{ij}, demonstrating the equivalence of these moduli and leaving an expression with just three independent moduli:

$$E = (c_{11}/2)(e_{xx}^2 + e_{yy}^2 + e_{zz}^2) + (c_{44}/2)(e_{yz}^2 + e_{zx}^2 + e_{xy}^2)$$
$$+ c_{12}(e_{yy}e_{zz} + e_{zz}e_{xx} + e_{xx}e_{yy}). \tag{10.7}$$

The bulk modulus, K, is obtained with $E = (1/2)K(\Delta V/V)^2$ by writing $e_{xx} = e_{yy} = e_{zz} = (1/3)\Delta V/V$ and equating the shear terms to zero,

$$K = (1/3)(c_{11} + 2c_{12}). \tag{10.8}$$

There are two independent shear moduli. For shears across any of the [100] crystal planes (normal to *x*, *y* or *z*) only the second term in Eq. (10.7) is relevant, so that

$$\mu_{100} = c_{44}. \tag{10.9}$$

The other shear modulus can be selected to be independent of c_{44} by taking a shear stress across a (110) plane (normal to a cube face diagonal). This gives

$$\mu_{110} = (1/2)(c_{11} - c_{12}). \tag{10.10}$$

The deformations leading to K, μ_{100} and μ_{110} are illustrated in Fig. 10.2. They relate more obviously to the moduli in Section 10.2 than c_{11}, c_{12} and c_{44} and can equally well be regarded as the three independent moduli of cubic crystals. The process of averaging them to obtain the moduli of an isotropic composite from the crystal constants is discussed in Section 10.4. In particular crystals the two shear moduli, represented by Eqs. (10.9) and (10.10), may happen to be nearly equal, making the crystals almost isotropic. The diamond structure is a special case. It is cubic, with three moduli, but Keating (1966) showed that there is a relationship between them, $2c_{44}(c_{11}+c_{12}) = (c_{11}-c_{12})(c_{11}+3c_{12})$, so that, although it is not isotropic, there are only two independent moduli.

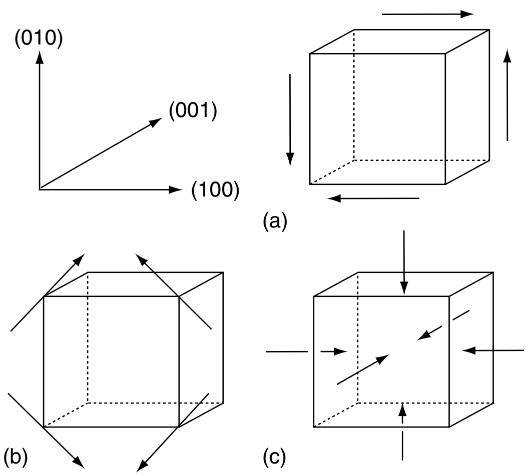


FIGURE 10.2 Stresses relative to the axes of a cubic crystal that cause strains represented by moduli (a) μ_{100} (b) μ_{110} (c) K, as given by Eqs. (10.8) to (10.10).

An observation from the elastic constants of common minerals with cubic structures is that the variations of P- and S-wave speeds with crystal orientation are generally negatively correlated. This accords with the simple minded notion that P-wave propagation in a 'hard', i.e. fast, direction is consistent with 'softer' transverse motions for S waves travelling in the same direction. However, the upper mantle anisotropy modelled by PREM gives faster horizontal waves for both P and S. In a layered structure both speeds are faster in the plane of layering but the effect is slight enough to make it an implausible explanation for the anisotropies of several per cent in the uppermost 200 km of the mantle in PREM. The major upper mantle minerals are olivine and pyroxenes, with orthorhombic structures (nine moduli), so it is evident that the simplistic idea that anisotropies for P and S waves would be opposite does not apply to these more complicated elasticities. Nonetheless, there is a feature of the wave propagation in cubic crystals that is relevant: for certain orientations, S-wave speed is very sensitive to polarization, i.e. direction of particle motion, and is consistent with the polarization splitting of S waves.

We refer in Section 10.1 to the elastic anisotropy of the uppermost part of the mantle, which is modelled as uniaxial or transversely isotropic in PREM. As a global average representation this is the form of anisotropy that must be expected even though transverse isotropy is unlikely to apply locally. The elastic symmetry of transverse isotropy coincides with that of hexagonal crystal structures, which have five moduli. They are identified with the speeds of waves parallel and perpendicular to the symmetry axis, that is vertical (radial) and horizontal. In the notation used by Dziewonski and Anderson (1981) in the presentation of PREM, they are *C*, *A* for vertical and horizontal P waves, *L* for S waves propagating radially and *L*, *N* for S waves propagating horizontally with vertical and horizontal polarizations respectively. A fifth modulus, *F*, is needed to represent wave speeds at intermediate angles, but is replaced in the PREM analysis with a dimensionless parameter $\eta = F/(A - 2L)$.

10.4 Relaxed and unrelaxed moduli of a composite material

As mentioned in Section 10.1, elastic isotropy normally results from a random alignment of crystals which are individually anisotropic and this is a good approximation for many rocks. They are composites of minerals that may have quite different elasticities and we need to relate the composite moduli to the elasticities of the components. There is no exact general solution to this problem, but several useful approximations have appeared over the long history of attempts to solve it. Watt *et al.* (1976) presented a comprehensive review. Mathematically, it is similar to the calculation of dielectric constant or conductivity of a composite and for much of this history electromagnetic theory led the way.

The simplest assumption is that, when stress is applied to a composite, all components are equally strained and that the stress is a volume-weighted average of the stresses on the components. This gives the Voigt formula, which, for the case of bulk modulus, K, is

$$K_V = (V_1K_1 + V_2K_2 + \cdots)/(V_1 + V_2 + \cdots), \qquad (10.11)$$

where V_1 etc. are volumes of the constituents. The assumption that the components are differently stressed implies a mismatch of stresses across boundaries between grains and cannot be valid in any real situation, but the Voigt approximation is taken as one extreme limit of the elasticity of a composite. The other extreme assumption is that all components are equally stressed, so that the total strain is a volume average of the constituent strains. This yields the Reuss limit, K_R, for the composite modulus,

$$1/K_R = (V_1/K_1 + V_2/K_2 + \cdots)/(V_1 + V_2 + \cdots). \qquad (10.12)$$

For small strains this is as unrealistic as the Voigt limit, but the two are extremes between which the truth must lie, and Hill (1952) argued that an average of the two limits,

$$K_{VRH} = (K_V + K_R)/2, \qquad (10.13)$$

normally gives an excellent approximation. This is known as the VRH (Voigt–Reuss–Hill) average elasticity. Sometimes geometric or harmonic means are used, or an average of two of these, but if something better than Eq. (10.13) is required there are alternative more restrictive limits that can be applied, as mentioned below.

This averaging procedure is appropriate for small stresses. In considering the response to a very high stress, or a very prolonged stress, the Reuss limiting elasticity (Eq. 10.12) must be used. In the deep Earth, where pressure greatly exceeds material strength, individual grains are deformed as necessary to equalize their ambient pressures, making the Reuss modulus, K_R, appropriate to the variation of compression with depth in the Earth. However, the small superimposed stresses in a seismic wave do not deform grains in this way and are better represented by Eq. (10.13). This introduces a complication to the seismological interpretation of the Earth's density profile (Chapter 17) and to the interpretation of internal properties by finite strain theory (Chapter 18), because the adiabatic bulk modulus obtained from seismic wave propagation is slightly larger than the modulus describing density variations in an adiabatic layer of mixed mineralogy.

Limits to the elastic moduli of a composite that are more restrictive than those of Voigt and Reuss can be applied if certain assumptions regarding grain geometry are justified (for details see Watt *et al.*, 1976). For the case of two constituents, A and B, the limits can be taken as the elasticities of media comprising spheres of A in a matrix of B and spheres of B in a matrix of A. This is the same as the dielectric problem in the classic work of J. C. Maxwell and leads to elastic limits explored by Hashin and Shtrikman (1963). However, Eqs. (10.11), (10.12) and (10.13) are adequate in most situations and are much easier to apply, especially if more than two constituents are involved or if finite strain theory must be used, as in Section 18.7.

10.5 Anelasticity and the damping of elastic waves

In the first paragraph of this chapter we mention inelastic deformation as a departure from elasticity observed when a temporary application of stress leaves a permanent effect. The following section and Chapter 11 pursue different aspects of this. Here we examine another effect that is subtly different and is termed anelasticity, to distinguish it from inelasticity, although the two do not always have different causes. Anelasticity is also a consequence of a departure from ideal elasticity (Hooke's law) but no permanent deformation is involved. It is observed when stress is cycled and strain follows with a delay. The result is a conversion of strain energy to heat in cyclically strained material. Its particular significance to geophysics is that it causes attenuation of seismic waves. The subject is reviewed by Knopoff (1964).

There are several processes, collectively termed internal friction, that contribute to anelasticity. It is useful to refer to two different types as linear and non-linear. Although they are not always clearly separable, they have different effects. In most rocks a linear range is observed for strain amplitudes less than about 10^{-6} and for greater strains there is an increasing non-linear component. If sinusoidal stress is imposed, linear internal friction is characterized by elliptical loops in a stress–strain diagram (Fig. 10.3), that is strain follows stress with a phase delay, ϕ. It results from relaxation phenomena, so that there is a frequency dependence, which, in the simplest situation of a single relaxation time, τ, gives a maximum ϕ at frequency $(1/2\pi\tau)$. This can be modelled by a Kelvin–Voigt spring and dashpot system (Fig. 10.4b). Linear internal friction is, by definition, amplitude independent, which means that φ is independent of stress or strain amplitude, as measured by loop length. Then loop shape (width/length) is constant at fixed frequency. The energy dissipation per cycle is the area of a loop in the stress–strain diagram, and so is a constant fraction of the peak strain energy.

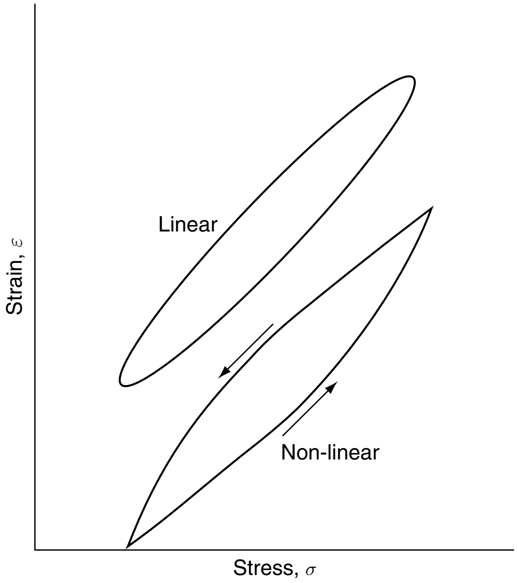


FIGURE 10.3 Forms of the stress–strain curve for a rock subjected to a strain cycle of amplitude less than 10^{-6} (linear) and much greater than 10^{-6} (non-linear). The loops are not to scale and the widths are greatly exaggerated to show the shape difference.

Although there is no sharp threshold for the onset of non-linear internal friction, it occurs at strain amplitudes that are large enough to cause grain boundary movements at the atomic level. This is a true hysteresis phenomenon in which the lag in strain, relative to stress, is determined by the level of stress rather than a time delay. When it is dominant it gives pointed stress–strain loops, as in the lower curve of Fig. 10.3. The loop shape varies with strain amplitude, becoming relatively wider as strain increases, so that the fractional energy dissipation per cycle increases with strain amplitude. That is the meaning of the term non-linear in this context. The local strain release in earthquakes is normally at least 10^{-5} and may be 10^{-4}, which is clearly in the range of non-linear anelasticity. Thus, the initial attenuation of seismic waves is non-linear and completely linear attenuation takes over only when waves have travelled a distance of order 10 times the smallest dimension of the faulted surface.

The effect of anelasticity on seismic waves is discussed in terms of the parameter Q, which is related to the fractional loss of energy per cycle

$$2\pi/Q = (-\Delta E/E). \tag{10.14}$$

For a simple linear relaxation process, with a single relaxation time, τ, Q is independent of strain amplitude but varies with frequency ($\omega/2\pi$) as

$$1/Q = \omega\tau/[1+(\omega\tau)^2], \tag{10.15}$$

which, as mentioned, gives peak dissipation at $\omega\tau = 1$. Special experimental conditions are required for the observation of isolated dissipation peaks with the form of Eq. (10.15) because, in a material such as rock, there is a superposition of many processes with different relaxation times, smearing the peaks into a continuum. It is often assumed that the total Q^{-1} is independent of frequency, as for non-linear anelasticity, although this is never quite true. It may appear that this makes the distinction between linear and non-linear anelasticity inconsequential, but there is an interesting difference in the effect on seismic waves.

For linear processes several waves may be superimposed, propagating and attenuating independently, so that a complex waveform may be treated as a sum of its independently decaying Fourier components. Within the assumption that frequency dependence of Q can be neglected, this allows Q to be estimated by a spectral ratio method of observing the greater attenuation of high frequencies with distance (or time). Non-linear processes may be truly hysteretic, that is the stress–strain hysteresis curve is followed independently of frequency and in this case a waveform propagates with attenuation but little variation in spectral content. Near to their sources seismic waves have strain amplitudes in the non-linear range and there is a transition to the more familiar linear regime only at distances of order ten times the dimensions of a fault face across which a displacement occurs.

There are alternative representations of the damping of waves or oscillations that can be related to Eq. (10.14). Writing this equation in differential form for a wave of period τ,

$$\frac{1}{E}\frac{\mathrm{d}E}{\mathrm{d}t} = -\frac{2\pi}{Q\tau}, \tag{10.16}$$

and integrating, we have

$$E = E_0 \exp\left(-\frac{2\pi}{Q\tau}t\right). \tag{10.17}$$

Since wave energy E is related to amplitude A as $E \propto A^2$,

$$A = A_0 \exp\left(-\frac{\pi}{Q\tau}t\right), \tag{10.18}$$

where t is the travel time. If the distance travelled in time t is x and the wavelength is λ, then

$$t/\tau = x/\lambda, \tag{10.19}$$

so that

$$A = A_0 \exp\left(-\frac{\pi}{Q\lambda}\cdot x\right) = A_0 \exp(-\alpha x), \tag{10.20}$$

where α is called the attenuation coefficient. In observations of oscillations a quantity called logarithmic decrement, δ, is often used.

$$\delta = -\frac{\mathrm{d}\ln A(n)}{\mathrm{d}n} = \frac{\pi}{Q} \approx -\frac{\Delta A}{A}, \tag{10.21}$$

which is the fractional loss of amplitude per cycle, n being the number of cycles. Not only is it convenient in many situations to plot or compute $\ln A$ vs n, but the linearity of this relationship is a direct test for linearity of the attenuation mechanism.

For a simple oscillatory system of resonant frequency f_0, Q, as used in these equations, is formally equivalent to the definition in terms of the width of a spectral line. The amplitude of forced vibration at frequency f depends on the difference between f and f_0, being greatest at $f = f_0$. Then if Δf is the frequency difference between the half power points on either side of resonance, that is the difference between the frequencies at which the vibration amplitude is reduced to $1/\sqrt{2}$ of the resonant value,

$$Q = f_0/\Delta f. \tag{10.22}$$

This equation can be applied also to the spectrum of a freely decaying oscillation, such as the Earth's free oscillations. In this case care is required either to avoid spectral lines that are significantly broadened by frequency splitting or to estimate the widths of the individual components of the lines. In Section 16.3 effects of lateral heterogeneity in the mantle are considered. This heterogeneity may slightly broaden the spectral lines and partly account for the lower Q values for free oscillations, relative to body waves.

For body waves Q may be estimated from wave decay using Eq. (10.18) or (10.20), but the corrections for geometrical spreading introduce troublesome uncertainties. These can be avoided by making a comparison of attenuations at different frequencies if the frequency dependence of Q is known. In this spectral ratio method it has normally been assumed that Q is independent of frequency over the range of interest. Then, substituting frequency $f = 1/\tau$ in Eq. (10.18) and differentiating with respect to f,

$$\frac{\mathrm{d}\ln(A/A_0)}{\mathrm{d}f} = -\frac{\pi}{Q}t. \qquad (10.23)$$

Thus, if a particular wave can be recorded at two points on its path and both records Fourier analysed, Q for the path difference may be estimated from frequency dependence of the ratio of spectral amplitudes at the two points. Q may be variable over the path, or there may be two different wave paths involved, in which case it is the difference in the quantity $\int Q^{-1}\mathrm{d}t$ for the two paths that is determined.

The reliability of the usual spectral ratio method is compromised by the assumption of frequency-independent Q. Laboratory observations of Q in rocks with strain amplitudes in the non-linear range have generally indicated frequency-independent Q. However, measurements made in the low strain, linear range that is more relevant to seismology, give Q increasing with frequency f as f^{ε}, with $\varepsilon \approx 0.3$. We can consider the magnitude of the consequent error in Eq. (10.23). When frequency-dependence of Q is allowed, the right-hand side of this equation is multiplied by the approximate factor $(1 - \mathrm{d}\ln Q/\mathrm{d}\ln f) = (1 - \varepsilon)$. Through most of the Earth there is a general increase in Q with frequency, so that this factor is less than unity and spectral ratio observations overestimate Q. But the error can hardly exceed a factor of 1.4 as a worst case and Q is generally so poorly known that this cannot be considered serious.

Q for P waves (Q_P) is systematically higher than for S waves (Q_S). It is the shear component of strain that is the principal cause of dissipation, even of P waves, and if pure compression were completely loss-less, Q_P/Q_S would be equal to the ratio of total strain energy to shear strain energy in a compressional wave. This is therefore the upper limit,

$$\frac{Q_P}{Q_S} \le \frac{K + 4\mu/3}{4\mu/3} = \frac{3(1-\nu)}{2(1-2\nu)}, \qquad (10.24)$$

which has numerical values between 2.27 and 2.67 for the PREM model of the mantle (Appendix F). Observed values of Q_P/Q_S are close to 2.0. Since P waves travel faster than S waves the ratio of attenuation coefficients, α_P, α_S, as in Eq. (10.20), is

$$\frac{\alpha_P}{\alpha_S} = \frac{Q_S}{Q_P} \cdot \frac{V_S}{V_P} \approx 0.27, \qquad (10.25)$$

where V_S,V_P are the wave speeds.

The frequency dependence of Q is not well constrained by observations. Body wave observations give higher values than are obtained from free oscillations and on this basis we might expect a still lower value at the 12-hour tidal period, for which Ray *et al.* (2001) reported a global average value of 280. This is comparable to the seismological value for the shear wave Q of the mantle, as listed in PREM (Dziewonski and Anderson, 1981). The frequency dependence may also be confused with scattering. Heterogeneities scatter wave energy according to the ratio of wavelength to the scale of the heterogeneities. This is especially important in the upper crust (Section 16.3). Another light on the frequency dependence of Q is an observation of its temperature variation in polycrystalline olivine samples by Jackson *et al.* (2005), who reported a strong temperature dependence above 900 °C at low frequencies (<1 Hz) but not in ultrasonic measurements in the megahertz range. This indicates a high-temperature onset of a viscoelastic effect.

10.6 Inelasticity, creep and flow

Creep is the progressive deformation of a material that remains coherent (unbroken) and

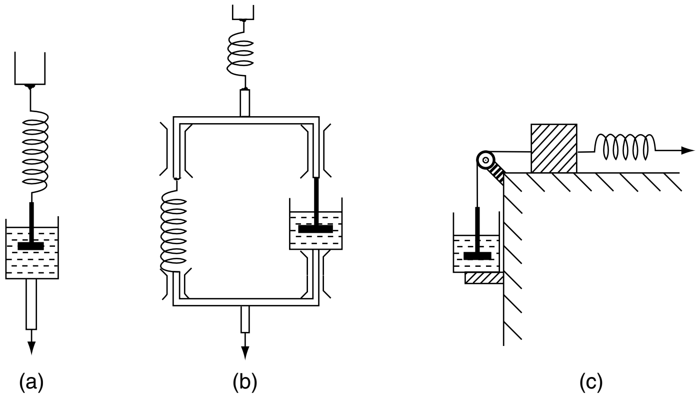


FIGURE 10.4 Mechanical models for (a) Maxwell (elastico-viscous), (b) Kelvin–Voigt (firmo-viscous) and (c) Bingham (plastico-viscous) solids. Firmo-viscosity is often represented without the immediate elastic response of the top spring. Note that to move the block in (c) friction must be overcome, and this represents a finite yield point.

homogeneous. It is one type of inelasticity. The other type is observed when a body has or develops localized weakness with concentrated deformation or fracture. This is subject to a rock mechanics approach (Chapter 11).

Several physical processes contribute to creep. Historically they have been represented by simple mechanical analogues that combine the elasticity of ideal springs with the viscous behaviour of dashpots (Fig. 10.4). These empirical models can be made to match a wide range of behaviour, especially if they are combined. The Maxwell model (Fig. 10.4a) represents steady state creep or progressive inelastic deformation superimposed on elastic strain. The deformation continues as long as the stress is applied and when it is removed only the elastic response is recovered. The convective deformation of the mantle is a process that can be represented by a Maxwell model.

The model of Kelvin and Voigt (Fig. 10.4b) can describe a deformation that saturates to a fixed limit with steady stress, due to work-hardening in a metallurgical situation. It is a relaxation model and is of geophysical interest as a representation of anelasticity, responsible for the attenuation of seismic waves (Section 10.5). It illustrates the distinction between relaxed and unrelaxed moduli. For oscillatory stresses that are very rapid compared with the relaxation time, τ, there is no appreciable response by the lower spring and dashpot and the response of the upper spring represents the unrelaxed modulus. For very slow stress cycles the lower spring follows the stress, giving an added response and a smaller, relaxed modulus. There is negligible hysteresis at either of these extremes, but at intermediate frequencies there is a partial response by the lower spring–dashpot combination and a phase lag between stress and strain, causing energy dissipation. This illustrates the relationship between the attenuation of seismic body waves and their velocity dispersion (frequency dependence of wave speed), although attenuation is not explicable by a single relaxation time but requires the superposition of many different relaxation times. The Reuss modulus of a composite (Eq. (10.12)) is an example of a relaxed modulus, whereas the Hill average modulus (Eq. (10.13)) is unrelaxed.

Departures from ideal elasticity in crystals are attributed to imperfections in crystal structure. The simplest of these are point defects, that is lattice vacancies and interstitial atoms (extra atoms squeezed into a lattice), that may move in response to an applied stress to positions that have lower energy in a stressed crystal. Also very important to inelastic and anelastic effects are extended linear defects, termed dislocations. A dislocation may be visualized as the mutual displacement of atoms on opposite sides of a cut extending through a crystal (Fig. 10.5). All of these defects are mobile, that is they move through a crystal more readily than atoms in a perfectly regular region. If a particular atomic displacement contributes to the deformation of a crystal in the sense of a stress imposed on it, then the stress makes that displacement more likely to occur. The displacements are impeded by interactions between defects and generally

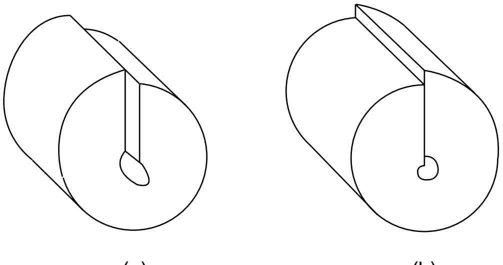


FIGURE 10.5 Displacements that produce (a) a screw dislocation and (b) an edge dislocation. The cylinders have been drawn as hollow to avoid difficulty with the discontinuity in displacement on the dislocation axes, which are cylinder axes.

depend on thermal activation, the probability of 'climbing' a barrier of energy E being proportional to $\exp(-E/kT)$ at temperature T. The effect of stress may be to reduce E for displacements in favoured directions and increase it for opposite movements, or merely give greater opportunity for favoured displacements, imparting a statistical bias to the movements that occur. Of course, macroscopic crystal deformation involves sequences of atomic movements with different barrier energies; the highest barriers are rate-controlling and determine the activation energy for the whole process. In fine grained materials, grain boundary effects become important. Although superficially similar to frictional sliding between grains, they involve individual atomic displacements that accommodate boundary fitting. Like diffusion or dislocation creep, they are relaxation phenomena.

Pressure, P, makes all atomic movements more difficult and so increases barrier energies. This is often represented by writing $E = E + PV^*$, with the constant V^* referred to as activation volume. However, in a geophysical context it is more convenient to relate E to melting point, T_M, by writing

$$E = gkT_M, \qquad (10.26)$$

where k is Boltzmann's constant and g is a dimensionless factor that is found to average about 27 for common minerals (Poirier, 2000). Equation (10.26) incorporates the pressure-dependence of E because melting depends on the same sort of atomic displacements as are responsible for solid creep. As discussed in Section 17.4, melting can be regarded as a free proliferation of crystal dislocations, a liquid being a crystal saturated with dislocations. The increase in T_M with pressure mirrors the increase in E, so that the variation of inelastic deformation with depth in the Earth is modelled with a theory of the variation of T_M with P.

Progressive deformation of the mantle occurs by steady state creep and can be represented by a general equation (Weertman and Weertman, 1992)

$$d\varepsilon/dt \equiv \dot{\varepsilon} = B(\sigma/\mu)^n \exp(-gT_M/T), \qquad (10.27)$$

where B is a constant. This equation incorporates both temperature and pressure dependences, by virtue of the pressure effect on T_M, and allows for a variety of mechanisms by the arbitrary index, n. The familiar case of Newtonian viscosity requires $n = 1$, with strain rate, $\dot{\varepsilon}$, proportional to stress, σ. Crystal deformation by diffusion of point defects (Nabarro–Herring creep) is of this form. Dislocation-dominated creep occurs in different regimes with different values of n, between 1 and 6; $n = 3$ is commonly favoured.

Regarding a liquid as a fully dislocated solid, we can consider fluid flow to be a special case of dislocation-mediated steady-state creep, with no essential distinction between solid and liquid, except for the concentration of dislocations. In this case, with $n = 1$ in Eq. (10.27), viscosity, η, is an unambiguous material property,

$$\eta = \sigma/\dot{\varepsilon}. \qquad (10.28)$$

Its SI unit is the pascal-second (Pa s). While validity of the $n = 1$ assumption is not as clear in the Earth, we nevertheless refer to viscosity for all layers below about 70 km, where the temperature is high enough for steady state creep to occur. The uppermost 70 km (the lithosphere) is too cool to be treated in this way and, depending on the phenomenon to be analysed, may be regarded as elastic (or perhaps subject to transient creep) or, for the uppermost layer, with deformation analysed by a rock mechanics approach, as in Chapter 11.

The homologous temperature, T/T_M, appears in the exponential term of the general creep law

(Eq. (10.27)) and rheological properties are very sensitive to it. For materials such as minerals, steady creep is observed only if T/T_M is reasonably high and at lower temperatures brittle failure occurs instead. In the Earth the distinction is apparent in the distribution of earthquakes, which are restricted to relatively cool regions. From a study of the thermal structure of the lithosphere at the point of subduction, McKenzie *et al.* (2005) reached the conclusion that earthquakes occur only where the temperature is below 600 °C. They interpreted this as the temperature distinguishing brittle failure from creep. From a detailed study of Californian earthquakes, Bonner *et al.* (2003) concluded that the limit is nearer to 400 °C. Taking the solidus temperature of the uppermost mantle as $T_M \approx 1400\,\mathrm{K}$, we can refer to a homologous temperature, $T/T_M = 0.5$ to 0.6, as the critical condition distinguishing seismic from aseismic behaviour. Although T_M is not well defined, because the different minerals have different melting points, we take this condition as an approximate guide to the variation in homologous temperature throughout the mantle. $T/T_M < \sim 0.5$ only near to the surface and in subducting slabs in the upper mantle and not at all in the lower mantle. Implications for lower mantle temperatures are considered in Section 19.5.

10.7 Frequency dependent elasticity and the dispersion of body waves

Consider the special case of a seismic pulse that starts as a δ function, a square pulse of infinitesimal duration. Its Fourier spectrum is initially white, that is, wave components of all frequencies have the same energy per unit frequency interval. Crests of all of the component waves coincide at the initiating δ function, but cancel elsewhere. But this cancellation fails as soon as the mix of components changes by selective removal of high frequencies. If we make the assumption that all frequencies travel at the same speed, then the coinciding crests stay together as a peak of the pulse. But it does not remain as a sharp δ function; it spreads out symmetrically to both earlier and later times. The peak would arrive at any point down the path at the time expected from the wave speed, so half of the pulse arrives faster than the wave speed. Worse, since there is no sharp onset of the pulse, it begins to arrive even before it has been initiated. This violation of the principle of causality demonstrates that there is an erroneous assumption: wave components of different frequencies cannot travel at the same speed. A frequency-dependent attenuation coefficient necessarily implies also frequency-dependent speed or wave dispersion. In the special case of frequency independent attenuation, that is $Q \propto \omega$, there would be no dispersion because all harmonic components of a pulse would be similarly attenuated and it would propagate undeformed, but with diminishing amplitude.

The problem of attenuation-related dispersion is discussed by Aki and Richards (2002) and the mathematical theory, based on the requirement of causality, is reviewed by Brennan and Smylie (1981). Most discussions emphasize the approximation of constant Q, that is Q independent of frequency, which yields a relationship between the phase speeds, v, at two frequencies and the Q of the medium,

$$\frac{v(f_1)}{v(f_2)} = 1 + \frac{1}{\pi Q}\ln\left(\frac{f_1}{f_2}\right). \tag{10.29}$$

This is a valuable result, showing that, for values of Q observed for the Earth, the dispersion is slight but not insignificant over the frequency range of seismology. It is much more important in comparing observations at seismic frequencies with laboratory measurements at MHz or GHz frequencies. In any case, Eq. (10.29) assumes constant Q, and so is only a rough approximation to the real Earth situation.

A more general treatment requires solution of an integral involving the attenuation coefficient (α in Eq. (10.20)). The variation in phase speed, v, with frequency, $\omega_0 / 2\pi$, can be obtained for an arbitrary variation of α with frequency, using alternative expressions that relate v to speed c at infinite frequency (Brennan and Smylie, 1981),

$$\frac{1}{v(\omega_0)} = \frac{1}{c} - \frac{1}{\pi\omega_0}\mathrm{pv}\int_{-\infty}^{\infty}\frac{\alpha(\omega)}{\omega_0-\omega}\mathrm{d}\omega$$
$$= \frac{1}{c} - \frac{2}{\pi}\mathrm{pv}\int_{0}^{\infty}\frac{\alpha(\omega)}{\omega_0^2-\omega^2}\mathrm{d}\omega, \qquad (10.30)$$

where pv indicates the Cauchy principal value of the integral and care is required in handling the singularities. The first version of this integral is the Hilbert transform of $\alpha(\omega)$. Provided $\alpha(\omega)$ is known over a range extending well beyond the frequency range of immediate interest, the high-frequency limits of the integrals present no problem because we are interested only in the differences in v between finite values of ω_0 and so subtract integrals from one another. A simple qualitative argument shows what result to expect. If α is independent of ω, corresponding to $Q \propto \omega$ or f, then the integrals in Eq. (10.30) are independent of ω_0 and there is no dispersion. In this case we can see that all of the Fourier components of the δ function pulse considered above are similarly attenuated. The pulse propagates with diminishing amplitude but no change in shape and no dispersion.

Now we can apply a simple amendment to Eq. (10.29) to allow for frequency dependent Q. If we write

$$Q \propto f^{\varepsilon} \qquad (10.31)$$

and consider an average of Q^{-1} over a moderate frequency range of interest, then the equation

$$\frac{v(f_1)}{v(f_2)} \approx 1 + \frac{(1-\varepsilon)}{\pi}\ln\left(\frac{f_1}{f_2}\right)\langle Q^{-1}\rangle \qquad (10.32)$$

satisfies the two conditions, $\varepsilon = 0$ (Eq. (10.29)) and $\varepsilon = 1$ (no dispersion). Equation (10.32) agrees with laboratory data on diverse materials, including plastics, metals, minerals and rocks, but it is an empirical approximation, justified by the fact that it is intermediate between the exact solutions of Eq. (10.30) for $\varepsilon = 0$ and $\varepsilon = 1$. Over the frequency range from body waves to free oscillations of the Earth, $\varepsilon \approx 1/3$ appears to be a reasonable approximation, so that dispersion is only about 2/3 as strong as in Eq. (10.29).

Linear anelastic mechanisms are relaxation phenomena, so that, after a disturbance to any element of material, causing an immediate elastic response, there is a further, delayed, anelastic response, which is an exponential relaxation towards an equilibrium state. Thermally activated movements of crystal dislocations past potential barriers and the redistribution of heat between compressed or dilated crystals with different thermodynamic properties are examples of processes that cause such relaxation. It follows that the elastic modulus has a higher value at high frequencies, which do not allow time for the relaxation to occur, than at lower frequencies. These are causes of a difference between relaxed and unrelaxed moduli, considered in the special case of composite materials in Section 10.4.

The variation of elastic modulus with frequency is twice as strong as the variation in phase speed, which depends upon the square root of elastic modulus. In some analyses what is here referred to as the elastic modulus is considered to be the real part of a complex modulus, with the imaginary part representing strain that is $\pi/2$ out of phase with stress. The real component, or modulus, is the gradient of the axis of a hysteresis loop, as in Fig. 10.3, and experiments confirm that the modulus increases with frequency. If the phase lag between strain and stress is δ then $Q^{-1} = \tan\delta \approx \delta$, which is the ratio of imaginary to real components of the modulus (not to be confused with the logarithmic decrement (Eq. (10.21)).

Since body waves have a slight positive dispersion (phase speed increasing with frequency), group speed is slightly greater than phase speed (see Eqs. (16.49) to (16.51)). Thus, at any particular frequency,

$$u = v\Big/\left(1 - \frac{1-\varepsilon}{\pi Q}\right). \qquad (10.33)$$

However, u is less than the limiting phase speed at high frequency, which corresponds to the value of v for the completely unrelaxed modulus.

It should be emphasized that the theory of body wave dispersion is valid only for linear attenuation mechanisms that operate at large distances from earthquakes. As we have mentioned, anelasticity may be non-linear within about ten fault dimensions. We have referred also to the observational problem that high frequency waves are more readily scattered than low frequencies and effects of scattering can be confused with attenuation.

11

Deformation of the crust: rock mechanics

11.1 Preamble

The theory of elasticity in Chapter 10, even the discussion of inelastic deformation, does not provide a description of the processes that led to the geological features that we see in the surface layers of the crust. We need a different theoretical approach to the deformation of materials such as crustal rocks that are granular, cool and at low pressure. What determines how or why rocks deform or break? How are the orientations of faults related to the stresses that cause them? How is crustal stress estimated? How is the safety of mines and tunnels assessed? Can laboratory observations of rock failure be extrapolated to describe earthquakes? These are questions addressed by the discipline of rock mechanics. In its modern form it is used mainly by mining engineers and the applied mathematicians who work with them. Here it is applied to tectonics.

The mathematical methodology of the subject uses the tensor notation. This is given a brief introduction in Section 10.3, in the discussion of elasticity of crystals, and is extended in this chapter. It is a convenient representation of stress patterns involving superimposed compression and shear, including rotation of reference axes, to identify principal stresses. The most obvious and useful application is to the criteria for mechanical failure.

Earthquakes are dramatic instances of failure and most of them are shallow, by which we mean that they are confined to the lithosphere, the cool surface layer of the Earth. The exceptions occur in the zones of subduction or convective downwelling, where planes of earthquake foci trace the penetration of the mantle by the cool lithospheric material to depths as great as 700 km. They are examples of mechanical failure, in which material loses its coherence. It is essentially a phenomenon of 'cool' material, at homologous temperatures $T/T_M < \sim 0.5$, where T_M is the melting point or solidus temperature. Most of the mantle is at higher temperatures and its convective deformation, described as creep in Section 10.6, occurs without sudden, localized failure.

Rock mechanics provides an interpretation of different styles of faulting in terms of stress orientations in the remote past as well as now. Earthquake fault plane solutions (Section 14.4) give the same information much more widely, but only at the present time. Stresses in aseismic areas are determined by application of rock mechanics equations to borehole measurements. All of these observations are combined to produce stress maps, which now cover a large fraction of the globe.

11.2 The tensor representation of stress and strain

Stress can be represented in three dimensions as a tensor. We consider an orthogonal coordinate system (x_1, x_2, x_3), written x_i, being the Cartesian (x, y, z), and identify three areas each with its normal aligned with one of the axes. Each area is represented by a unit vector $\mathbf{n}_i$ in the direction of the normal and a magnitude A_i. In the convention of continuum mechanics, stress is the force per unit area exerted on material on the negative side of the area by material on the positive side. An area can be subjected to three independent forces, one normal to it (in the direction of $\mathbf{n}_i$), and two tangential. With three force components acting on each of the three areas, a total of nine stress components is possible. However, as we shall see, only six are independent. A normal stress is a force, parallel to the area normal, divided by the area on which it acts. It can be either extensional or compressional depending on whether the force is directed parallel or antiparallel to the area normal. In this convention an extensional stress is positive but the opposite sign convention is used in rock mechanics, in which compressions predominate. The convention that compression is positive is also used in treatments of finite strain and high pressure equations of state (Chapter 18). Recognizing the danger of confusion, we use the continuum mechanics convention where this is standard and convert to the rock mechanics convention for the discussion of the Mohr circle in Section 11.4 and crustal stresses in the following sections.

Shear stresses are the tangential forces per unit area. A component of the stress tensor, written σ_{ij}, corresponds to a force acting in direction i on unit area of a plane with normal j. Shear stress is positive if the force acts in the positive i direction on the positive side of the j plane. With the corresponding Cartesian notation,

$$[\sigma_{ij}] = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix} = \begin{bmatrix} \sigma_{xx} & \sigma_{xy} & \sigma_{xz} \\ \sigma_{yx} & \sigma_{yy} & \sigma_{yz} \\ \sigma_{zx} & \sigma_{zy} & \sigma_{zz} \end{bmatrix}. \quad (11.1)$$

In one dimension, strain is a fractional change in length. In three dimensions, six strains are possible, three length changes, in the x_1, x_2 and x_3 directions, denoted e_{11}, e_{22}, e_{33}, and three shears e_{12}, e_{23}, e_{13}. The shear strain is defined as half the angular change in the right angle defined in the material by the x_i, x_j axes prior to deformation. From this it is evident that $e_{ij} = e_{ji}$, and that there are only three independent shear strains. This is mentioned in Section 10.3 and explained again below.

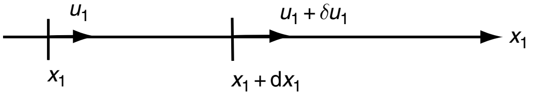


FIGURE 11.1 Variation of displacement u_1 in the direction x_1, giving the strain e_{11}.

After application of a stress, the new positions of material points in a body are given by the displacement field $\mathbf{u}(x,y,z)$. Strains are gradients in the displacement field. Consider a body subjected to a stress that causes a displacement deformation field $\mathbf{u} = (u_1,u_2,u_3)$. Let the displacement in the x_1 direction at two points a distance dx_1 apart be u_1 and $u_1 + du_1$ (Fig. 11.1). From a Taylor expansion, $u_1(x_1 + dx_1) = u_1(x_1) + (\partial u_1/\partial x_1)dx_1 + \cdots$, we retain only linear terms $du_1 = (\partial u_1/\partial x_1)dx_1$, that is we consider only linear or infinitesimal elasticity theory, applicable to small strains. The strain e_{11} is the change in length, $(\partial u_1/\partial x_1)dx_1$, per unit original length dx_1, that is

$$e_{11} = \partial u_1/\partial x_1. \quad (11.2)$$

We treat shear strains in a similar manner. Figure 11.2 shows the angle changes to what are right angles in unstrained material. Shear strain is

$$e_{12} = 1/2(\theta_1 + \theta_2) = 1/2(\partial u_2/\partial x_1 + \partial u_1/\partial x_2). \quad (11.3)$$

If $\theta_1 \neq \theta_2$ the body rotates, with rotation defined as $(\theta_1 - \theta_2)/2 = \frac{1}{2}\left(\frac{\partial u_2}{\partial x_1} - \frac{\partial u_1}{\partial x_2}\right)$.

We can generalize these results to three dimensions,

$$e_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right), \quad (11.4)$$

where i and j take on the values 1, 2 and 3.

Consider the volume $dx_1dx_2dx_3$ (Fig. 11.3), subjected to shear stresses in the x_1x_2 plane. The tangential forces acting on each side from the material external to the volume are $\sigma_{21}dx_2dx_3$, $-\sigma_{21}dx_2dx_3$, $\sigma_{12}dx_1dx_3$ and $-\sigma_{12}dx_1dx_3$. Then the torque is $\sigma_{12}dx_1dx_3.dx_2 - \sigma_{21}dx_2dx_3.dx_1 = (\sigma_{12} - \sigma_{21})dx_1dx_2dx_3$. For the torque to be

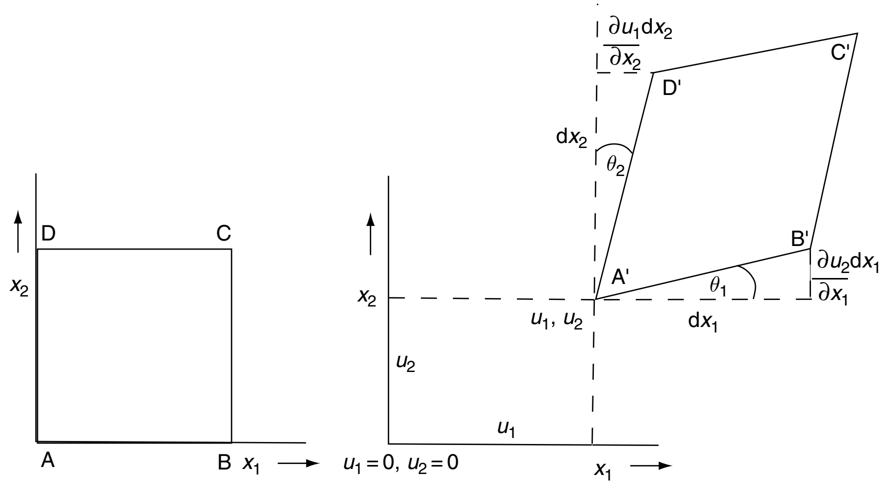


FIGURE 11.2 Displacements due to shear strain. Strains involve variations in displacements as a function of the coordinate system.

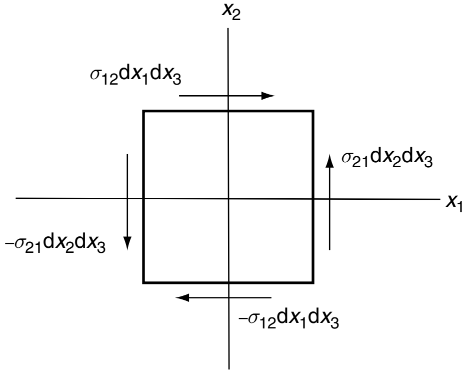


FIGURE 11.3 Forces on the faces of an elementary volume from a stress field σ_{ij}.

independent of volume, $dx_1dx_2dx_3$, $\sigma_{12} = \sigma_{21}$. Similarly $\sigma_{23} = \sigma_{32}$ and $\sigma_{13} = \sigma_{31}$. Generalizing, $\sigma_{ij} = \sigma_{ji}$, and, because in isotropic material shear stress is proportional to the corresponding strain, $e_{ij} = e_{ji}$, with the result that only six elements of the stress (strain) tensor are independent.

11.3 Hooke's law in three dimensions

For an isotropic material two independent moduli are required and there are several ways of choosing them (Sections 10.1 and 10.2). In most geophysics these are bulk modulus, K, relating pressure to volume change, and rigidity modulus, μ, relating shear stress to change of shape with constant volume. A different choice arises from the tensor notation in which the three dimensional version of Hooke's law is expressed as

$$\begin{aligned}
\sigma_{11} &= \lambda(e_{11} + e_{22} + e_{33}) + 2\mu e_{11}, \\
\sigma_{22} &= \lambda(e_{11} + e_{22} + e_{33}) + 2\mu e_{22}, \\
\sigma_{33} &= \lambda(e_{11} + e_{22} + e_{33}) + 2\mu e_{33}, \\
\sigma_{12} &= 2\mu e_{12}, \\
\sigma_{13} &= 2\mu e_{13}, \\
\sigma_{23} &= 2\mu e_{23},
\end{aligned} \tag{11.5}$$

where λ and μ are referred to as the Lamé coefficients but μ is familiar as the shear modulus. Equations (11.5) can be written in a convenient shorthand as

$$\sigma_{ij} = \lambda e \delta_{ij} + 2\mu e_{ij}, \tag{11.6}$$

where δ is the Kroeneker delta for which

$$\begin{aligned}
\delta_{ij} &= 1 \text{ if } i = j, \\
\delta_{ij} &= 0 \text{ if } i \neq j.
\end{aligned} \tag{11.7}$$

$e = e_{11} + e_{22} + e_{33}$ is the sum of fractional changes in length in three directions which gives the relative volume change. The relationships to the alternative isotropic moduli are here re-derived by application of Eq. (11.6). Pressure is defined as the average normal stress

$$p = -P = (\sigma_{11} + \sigma_{22} + \sigma_{33})/3, \tag{11.8}$$

where P has compression positive, as used in rock mechanics and interior stresses in the Earth. Bulk modulus, K, the ratio of pressure to volume change, is obtained by summing the first three of Eqs. (11.5):

$$K = P/e = (\lambda + 2/3\ \mu). \tag{11.9}$$

Poisson's ratio, ν, is not strictly a modulus, but a dimensionless ratio. However, it is a useful parameter that is always listed with elastic moduli. Let uniaxial stress σ_{11} be applied in the x_1 direction to a rectangular block of elastic material with the other sides free so that the other two stress components are zero. The block expands laterally if the body is compressed axially or will contract laterally if it is extended. Poisson's ratio is the negative ratio of lateral to longitudinal strains (a negative sign because the strains are of opposite signs). Equations (11.5) become

$$\begin{aligned}\sigma_{11} &= (\lambda + 2\mu)e_{11} + \lambda e_{22} + \lambda e_{33},\\ 0 &= \lambda e_{11} + (\lambda + 2\mu)e_{22} + \lambda e_{33},\\ 0 &= \lambda e_{11} + \lambda e_{22} + (\lambda + 2\mu)e_{33}.\end{aligned} \tag{11.10}$$

By symmetry, the lateral strains are equal, $e_{22} = e_{33}$, and Poisson's ratio is

$$\nu = -e_{22}/e_{11} = \lambda/(2\lambda + 2\mu). \tag{11.11}$$

An approximation that is sometimes used to simplify rock mechanics equations is $\lambda = \mu$, giving $\nu = 1/4$, referred to as a Poisson solid. While the simplification may be justified where rough solutions suffice, it underestimates Poisson's ratio for most solid materials, for which $\nu = 0.3$ is a better approximation at low pressure, and it increases systematically with pressure (for a reason discussed in Section 18.8). For a fluid, $\mu = 0$, so that by Eq. (11.11) Poisson's ratio is 0.5 and this applies to the outer core and oceans.

Young's modulus, E, is the ratio of longitudinal stress to strain, with no lateral stress, $\sigma_{22} = \sigma_{33} = 0$, as in the consideration of Poisson's ratio in Eq. (11.11).

$$E = \frac{\sigma_{11}}{e_{11}} = \frac{\mu(3\lambda + 2\mu)}{\lambda + \mu}. \tag{11.12}$$

We can use Eqs. (11.11) and (11.12) to solve for the Lamé coefficients, λ and μ, in terms of Young's modulus and Poisson's ratio:

$$\begin{aligned}\lambda &= \frac{E\nu}{(1+\nu)(1-2\nu)},\\ \mu &= \frac{E}{2(1+\nu)}.\end{aligned} \tag{11.13}$$

If a linear stress σ_{11} produces a strain e_{11} with lateral strain prevented, then the first of Eqs. (11.5) gives

$$\chi = \sigma_{11}/e_{11} = (\lambda + 2\mu). \tag{11.14}$$

This is the modulus controlling P-wave propagation in the Earth because the juxtaposition of compressed and extended layers ensures that $e_{22} = e_{33} = 0$. The lateral stresses

$$\sigma_{22} = \sigma_{33} = \frac{\lambda}{\lambda + 2\mu}\sigma_{11} = \frac{\nu}{1-\nu}\sigma_{11} \tag{11.15}$$

are self-adjusted to make this so. A more useful and familiar relationship for χ is obtained by substituting for λ by Eq. (11.9):

$$\chi = K + \frac{4}{3}\mu, \tag{11.16}$$

and we note that

$$\chi = (\lambda + 2\mu) = \frac{(1-\nu)E}{(1+\nu)(1-2\nu)}. \tag{11.17}$$

These relationships are summarized in Appendix D.

11.4 Tractions, principal stresses and rotation of axes

The stress tensor can be thought of as a table of coefficients that can be multiplied by an area vector to give the forces on that area. Consider an element of area of size A (Fig. 11.4) with direction cosines n_j. The forces per unit area applied to the material on the negative side of A by material on the positive side are called the tractions and are given by

$$T_i = \sum_{j=1}^{3} \sigma_{ij} n_j = \sigma_{ij} n_j \tag{11.18}$$

using the Einstein summation rule for repeated indices.

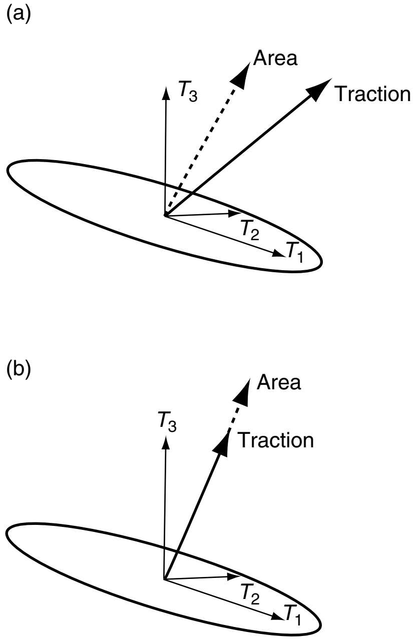


FIGURE 11.4 Tractions are forces applied by material on the positive side of an elementary area under stress. Case (a) is the general one and if the vector sum of the tractions is normal to the area, as in case (b), then the normal is a direction of principal stress.

Thus

$$\begin{aligned}T_1 &= \sigma_{11}n_1+\sigma_{12}n_2+\sigma_{13}n_3,\\ T_2 &= \sigma_{21}n_1+\sigma_{22}n_2+\sigma_{23}n_3,\\ T_3 &= \sigma_{31}n_1+\sigma_{32}n_2+\sigma_{33}n_3.\end{aligned} \tag{11.19}$$

For any stress field, σ_{ij}, in an isotropic material it is possible to find three orthogonal unit areas across which the tractions have no tangential components, that is vector **T** is parallel to the area vector **n** (Fig. 11.4(b)), which means that **n** is a direction of principal stress. This requires the traction components to be proportional to the direction cosines of the areas. Let the constant of proportionality be l, $T_i = l n_i$. Substituting Eq. (11.19) for T_i in Eq. (11.18) and using the Einstein summation rule, with the Kroenecker δ_{ij}, as in Eq. (11.7),

$$\sigma_{ij}n_j = l\delta_{ij}n_j,$$

or

$$\sigma_{ij}n_j - l\delta_{ij}n_j = 0.$$

To have a solution to this set of simultaneous equations the determinant must be zero,

$$\det(\sigma_{ij} - l\delta_{ij}) = \left|\sigma_{ij} - l\delta_{ij}\right| = 0. \tag{11.20}$$

Equation (11.20) is the characteristic equation for an eigenvalue problem. In three dimensions it is cubic in l,

$$\begin{vmatrix}\sigma_{xx}-l & \sigma_{xy} & \sigma_{xz}\\ \sigma_{yx} & \sigma_{yy}-l & \sigma_{yz}\\ \sigma_{zx} & \sigma_{zy} & \sigma_{zz}-l\end{vmatrix} = 0, \tag{11.21}$$

which has three roots $\{l_1, l_2, l_3\}$, corresponding to the three principal stresses. The corresponding eigenvectors are the three unit (vector) areas corresponding to each of l_1, l_2, l_3.

To find the eigenvector corresponding to l_1, for example, we require three equations in three unknowns $\{n_i\}$. Equation (11.21) comprises two independent equations. The third is the orthonormal property of the unit area vectors $\sum_{i=1}^{3} n_i^2 = 1$. Thus for l_1 the equations are

$$\begin{aligned}n_1^2 + n_2^2 + n_3^2 &= 1,\\ \sigma_{11}n_1 + \sigma_{12}n_2 + \sigma_{13}n_3 &= l_1 n_1,\\ \sigma_{21}n_1 + \sigma_{22}n_2 + \sigma_{23}n_3 &= l_1 n_2.\end{aligned} \tag{11.22}$$

An example in two dimensions

In two dimensions Eq. (11.22) reduces to

$$\begin{vmatrix}(\sigma_{11}-l) & \sigma_{12}\\ \sigma_{12} & (\sigma_{22}-l)\end{vmatrix} = 0.$$

The characteristic equation becomes

$$\begin{aligned}(\sigma_{11}-l)(\sigma_{22}-l) - \sigma_{12}^2 &= 0,\\ l^2 - l(\sigma_{11}+\sigma_{22}) + \sigma_{11}\sigma_{22} - \sigma_{12}^2 &= 0,\end{aligned} \tag{11.23}$$

the solution of which is

$$l = \frac{(\sigma_{11}+\sigma_{22})}{2} \pm \sqrt{\left[\frac{(\sigma_{11}-\sigma_{22})}{2}\right]^2 + \sigma_{12}^2}. \tag{11.24}$$

Suppose that the stress field is a pure shear ($\sigma_{11} = \sigma_{22} = 0$). Then $l = \pm\sigma_{12}$.

Choose $l_1 = +\sigma_{12}$. Solving for the eigenvector,

$$\begin{aligned}n_1^2 + n_2^2 &= 1,\\ \sigma_{12}n_2 &= l_1 n_1,\end{aligned}$$

giving $n_2 = n_1$, and therefore

$$n_1 = n_2 = 1/\sqrt{2}.$$

As is obvious in this simple case the principal stresses are a tension and compression equal to the value of the shear stress and oriented at 45° to the shear direction.

Rotation of stress or strain tensors

Suppose that we wish to determine the components of the stress tensor in a rotated coordinate system x'_i relative to an unrotated system x_i. The traction force vector in the rotated system is given by Eq. (11.19). The force components resolved normal and tangential to the axial planes of the rotated system give the new normal and shear stresses, respectively. For example, in two dimensions, if the axes are rotated anticlockwise by θ, the direction cosines of the area with normal x'_i are $n_j = [\cos\theta, \sin\theta]$. From Eq. (11.19) the tractions are $T_1 = (\sigma_{11}\cos\theta + \sigma_{12}\sin\theta)$; $T_2 = (\sigma_{21}\cos\theta + \sigma_{22}\sin\theta)$. Stresses in the rotated system, σ'_{11} and σ'_{12}, are found by resolving the traction vector normal and parallel to the area:

$$\begin{aligned}\sigma'_{11} &= (T_1\cos\theta + T_2\sin\theta);\\ \sigma'_{12} &= (-T_1\sin\theta + T_2\cos\theta).\end{aligned} \tag{11.25}$$

Inserting T_1 and T_2 we obtain

$$\begin{aligned}\sigma'_{11} &= \sigma_{11}\cos^2\theta + \sigma_{22}\sin^2\theta + \sigma_{12}\sin 2\theta,\\ \sigma'_{12} &= \frac{\sigma_{22}-\sigma_{11}}{2}\sin 2\theta + \sigma_{12}\cos 2\theta.\end{aligned} \tag{11.26}$$

Repeating for the area with normal x'_2, i.e., $\theta \to \frac{\pi}{2} + \theta$,

$$\sigma'_{22} = \sigma_{11}\sin^2\theta + \sigma_{22}\cos^2\theta - \sigma_{12}\sin 2\theta. \tag{11.27}$$

By examining (11.26) and (11.27) we see that the procedure for finding tractions in the rotated system, then resolving them to normal and tangential directions to find the stresses, can be expressed in matrix form as the product of two rotations,

$$R_1 = \begin{bmatrix}\cos\theta & \sin\theta\\ -\sin\theta & \cos\theta\end{bmatrix}, \quad R_2 = \begin{bmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{bmatrix}. \tag{11.28}$$

Then $\sigma' = R_1\sigma R_2 = R\sigma R^{\mathrm{T}}$, where T refers to the transpose of a matrix.

Generalizing to three dimensions, we let $R_{ij} =$ cos(angle between x_i and x'_j):

$$\sigma'_{ij} = \sum_{m=1}^{3}\sum_{n=1}^{3} R_{im}\sigma_{mn}R_{nj} = R\sigma R^{\mathrm{T}}. \tag{11.29}$$

The Mohr Circle

Consider a special case with the x_i coordinate axes so chosen that $\sigma_{12} = \sigma_{xy} = 0$, that is σ_{11} and σ_{22} are principal stresses, which we will denote by σ_1 and σ_2. After rotation by angle θ, Eqs. (11.26) and (11.27) may be rewritten as

$$\begin{bmatrix}\sigma'_{11} & \sigma'_{12}\\ \sigma'_{21} & \sigma'_{22}\end{bmatrix} = \begin{bmatrix}\frac{\sigma_1+\sigma_2}{2} - r\cos(2\theta) & r\sin(2\theta)\\ r\sin(2\theta) & \frac{\sigma_1+\sigma_2}{2} + r\cos(2\theta)\end{bmatrix}, \tag{11.30}$$

where

$$r = \frac{\sigma_2 - \sigma_1}{2}. \tag{11.31}$$

From Eq. (11.30) we see that the normal and shear stresses can be represented by a circle diagram (Fig. 11.5). The axes of the diagram are the normal stress and shear stress components (σ_n, τ). The centre of the circle is located at $\sigma_n = \frac{\sigma_1+\sigma_2}{2}, \tau = 0$. The radius of the circle is r (Eq. (11.31)). The stresses in a plane with its normal at angle θ to the direction of the x_1 axis are found by drawing the vector **r** from the point $\left(\frac{\sigma_1+\sigma_2}{2}, 0\right)$ at angle 2θ to the σ_n axis. The τ axis component gives the shear in the rotated system; the σ_n axis component gives one normal stress; the σ_n axis component of negative **r** gives the other normal stress.

To convert the circle diagram (Fig. 11.5) from the elastic (compressions negative) to the rock mechanics convention (compressions positive) the normal stresses must be reversed in sign. We rotate the circle diagram about the vertical axis, and change the sign of r in Eq. (11.31) to $r = \frac{\sigma_1 - \sigma_2}{2}$ (see Fig 11.6(a)). Then σ_1 is the largest (positive) compressive stress. The circle so formed is called the Mohr circle (Fig. 11.6(a))

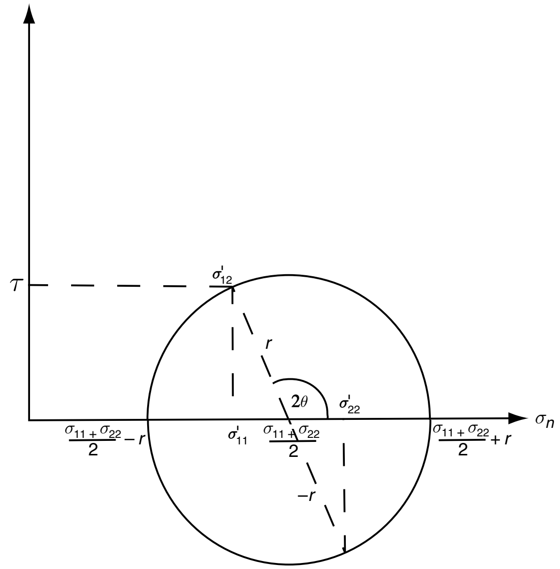

τ
σ'_{12}
r
2θ
σ'_{22}
σ_n
$\frac{\sigma_{11}+\sigma_{22}}{2}-r$
σ'_{11}
$\frac{\sigma_{11}+\sigma_{22}}{2}$
$-r$
$\frac{\sigma_{11}+\sigma_{22}}{2}+r$

FIGURE 11.5 Shear stress (vertical axis) plotted against normal stresses (horizontal axis). Shear stress in a plane at any angle to the *x*-axis is obtained by the projection of the **r** vector on the vertical axis. Normal stresses are obtained by the projections of the **r** vector and negative **r** vector on the horizontal axis.

after its inventor, O. Mohr. The intercepts of the Mohr circle with the horizontal axis are the principal stresses σ_1, σ_2. When the reference plane rotates by an angle θ in physical space, the corresponding point in the Mohr circle diagram (Fig. 11.6(a)) moves around the circle by an angle 2θ. From the Mohr circle we observe that the maximum shear occurs at $2\theta = \frac{\pi}{2}$, or $\theta = 45^\circ$, to the principal stress direction, and has a value equal to half the difference in magnitude of the maximum and minimum principal stresses. From Eq. (11.30) we have

$$\sigma'_{xx} = \frac{\sigma_1 + \sigma_2}{2} + \frac{\sigma_1 - \sigma_2}{2}\cos(2\theta),$$
$$\sigma'_{yy} = \frac{\sigma_1 + \sigma_2}{2} - \frac{\sigma_1 - \sigma_2}{2}\cos(2\theta),$$
$$\sigma'_{xy} = \tau = \frac{\sigma_2 - \sigma_1}{2}\sin(2\theta), \qquad (11.32)$$

from which we can derive the relation

$$\sigma'_{xx} - \sigma'_{yy} = (\sigma_1 - \sigma_2)\cos(2\theta) \qquad (11.33)$$

which we use in Chapter 13.

The three-dimensional case involves three principal stresses, σ_1, σ_2, σ_3. Let $\sigma_1 > \sigma_2 > \sigma_3$. We can apply the Mohr circle analysis on each plane containing two of the principal stresses, i.e., $\{\sigma_1,\sigma_2\}$, $\{\sigma_1,\sigma_3\}$, $\{\sigma_2,\sigma_3\}$, which results in three Mohr circles as in Fig. 11.6(b). Then each circle

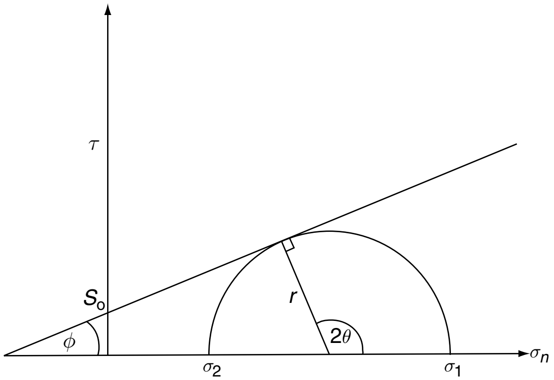


FIGURE 11.6(a) Mohr circle. Special case where the horizontal axis is the direction of greatest principal stress. The sloping line gives the frictional strength. Frictional failure occurs when the Mohr circle intersects the friction line.

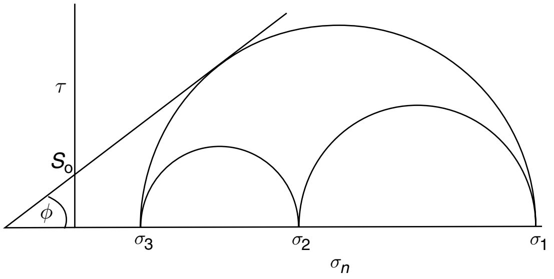


FIGURE 11.6(b) Mohr circle for three dimensions. Each circle represents a plane containing two of the principal stresses. The failure line occurs in the plane with the greatest shear stress.

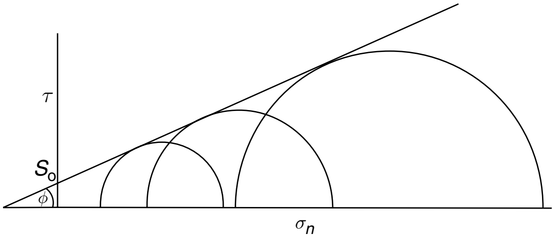


FIGURE 11.6(c) Increasing pressure causes stability against frictional failure by moving the centres of Mohr circles to the right, which means greater shears must be developed for their radii to reach the failure line.

describes stress components on a plane lying between the principal stress pair. We see that in Fig. 11.6(b) the maximum shear occurs in the plane containing the intermediate principal stress and bisects the angle between the greatest and least principal stresses.

11.5 Crustal stress and faulting

Since any stress pattern can be resolved into three principal stresses, σ_1, σ_2, σ_3, and deep in the Earth these are all large compared with the differences between them, it is often convenient to identify the average as a hydrostatic pressure, as in Eq. (11.8), recalling that $p=(\sigma_1+\sigma_2+\sigma_3)/3$, and we refer to the differences as deviatoric stresses. These are generally small enough to be treated by the linear (infinitesimal) elasticity theory, allowing a simple superposition on hydrostatic compression, even when the response to hydrostatic pressure is strong enough to require analysis by a finite strain theory, as in Chapter 18. Except near to the surface, the hydrostatic pressure, P, is approximated by the lithostatic or overburden pressure,

$$P \approx \int_0^z \rho g \, \mathrm{d}z, \tag{11.34}$$

which can be used as one of the equations needed for complete specification of the stress pattern. In common usage this is a positive quantity, but for substitution in the standard equations of elasticity theory, in which extension is positive, pressure is taken as the negative quantity $-P$. Then the tectonic stresses causing deformation, the deviatoric stresses, are obtained from the total stresses by subtracting the pressure, p:

$$\tau_{ij} = \sigma_{ij} - p\delta_{ij} = \sigma_{ij} + P\delta_{ij}. \tag{11.35}$$

Faults are often planes of weakness established by repeated movements, but their original causes were tectonic stresses in directions that determined the fault orientations. Tectonic movements, ultimately driven by mantle convection, persist in the same directions for many millions of years, so that faults reflect the driving forces that caused them, even where established planes of weakness may now have a strong influence. Similarly, faults that are no longer active indicate the orientations of stresses that moved them many years ago. This is the reason for interest in the classification of faults. They are of three types, illustrated in Fig. 11.7, corresponding to different stress orientations, as classified by Anderson (1905). The Anderson faulting criteria are summarized in Table 11.1.

Table 11.1 Anderson faulting criteria. σ_1, σ_2, σ_3, are maximum, intermediate and minimum principal stresses, respectively. Note that 'strike' means the orientation of the surface trace of the fault

	σ_1	σ_2	σ_3
Normal fault	vertical	horizontal along strike	horizontal
Strike slip fault	horizontal (angle to strike)	vertical	horizontal
Reverse (thrust) fault	horizontal	horizontal along strike	vertical

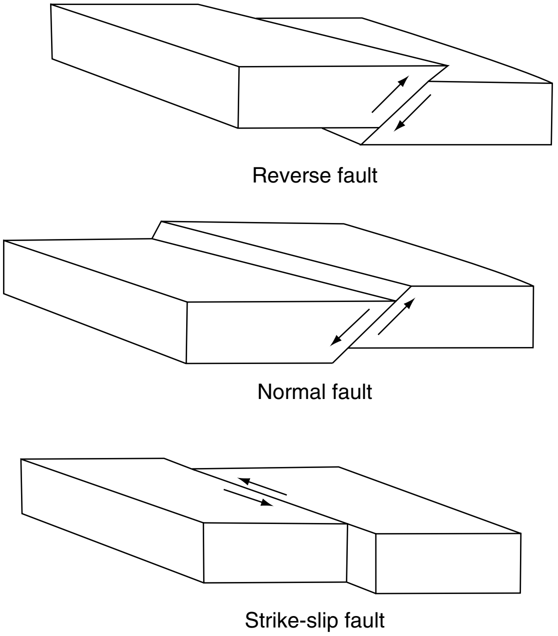


FIGURE 11.7 Different fault types.

The Anderson criteria are derived from two basic principles: (i) there are no stresses, normal or shear, across a free horizontal surface, and (ii) the largest shear stress occurs across planes between the directions of maximum and minimum principal stress. The existence of the free surface requires that one of the principal stress directions be vertical, normal to the surface. This applies not just to the surface itself but to all depths much smaller than the horizontal scale over which the stresses are averaged, and it is therefore a constraint on theories of the causes of geologically observed surface faults. Thus, the other two principal stresses are horizontal and there are only three ways in which they can be oriented. The vertical stress may be the direction of maximum, σ_1, intermediate, σ_2, or minimum principal stress, σ_3, and in each case the other two principal stresses are the horizontal stresses. As generally applied, the principal stresses are converted to the rock mechanics convention with compression positive. With this convention, the ranking $\sigma_1 > \sigma_2 > \sigma_3$ corresponds to the most compressive to least compressive stress.

In a region of compression, σ_1 and σ_2 are both horizontal, with σ_3 vertical, and thrust (reverse) faulting (Fig 11.7) occurs. Extension, in which both horizontal stresses are smaller than the vertical (and are negative at the surface), causes normal faulting (Fig. 11.7). Strike-slip faulting (Fig. 11.7) occurs when the maximum and minimum stresses are horizontal.

An unconfined rock will fail when subjected to shear stress greater than its internal strength, S_0. The failure usually takes the form of relative motion across a plane that is oriented in the direction of maximum shear, i.e. at 45° to the maximum and minimum principal stresses. At depth, rocks are more resistant to shear failure because they are subjected to confining pressure of the overburden. In addition to overcoming intrinsic internal strength, shears must be large enough to overcome frictional effects from the overburden pressure. The frictional stress is equal to the product of normal stress, σ_n, across a fault plane (with compression taken as positive) and a friction coefficient. The failure plane is then not the 45° plane of maximum shear, because the frictional stress is reduced by turning the normal to the failure plane towards the axis of minimum stress. This occurs at an angle such that the reduction in shear stress is compensated by a reduction in the normal stress. Coulomb's equation describes this process and gives the condition for failure as

$$\tau = \mu_f \sigma_n + S_0, \tag{11.36}$$

where τ is the limiting shear stress at which failure occurs, μ_f is the friction coefficient, and S_0 is the internal strength or cohesion. The straight line in Fig. 11.6(a) gives the failure criterion (Eq. (11.36)), which lies tangent to the limiting Mohr circle at the point of failure. The rock is stable for stress fields having circles that lie below the failure line. As Fig. 11.6(c) illustrates, increasing pressure stabilizes against failure because larger shears are required to generate Mohr circles that intersect the failure criterion.

From laboratory observations, Byerlee (1978) found that a value $\mu_f = 0.6$ to 0.85 applies to many common rock types. The lower value applies to wet rock or rock under high confining pressure and the upper value to dry rock at low confining pressure. Equation (11.36), known as Byerlee's law, is the equation of a straight line (a better approximation is a convex curve) that is plotted on the Mohr diagram to describe conditions for failure. At the point of failure, 2θ of Eq. (11.30) is an obtuse angle. Therefore, the angle, θ, between the normal to the plane of failure and maximum principal stress direction is greater than 45°. The slope of the failure line is $\tan\phi$, where ϕ is referred to as the angle of internal friction, so that

$$\mu_f = \tan\phi. \tag{11.37}$$

Then from Fig. 11.6(a),

$$2\theta = \frac{\pi}{2} + \phi. \tag{11.38}$$

In structural geology one is interested in δ, the angle between the dip of the fault plane and σ_1, which is the complement of θ. Then

$$\delta = \frac{\pi}{2} - \theta = \frac{1}{2}\tan^{-1}(1/\mu_f). \tag{11.39}$$

For $\mu_f = 0.85$, $\delta = 24.8°$. In general, failure planes are predicted to occur at small angles to the directions of maximum principal stress.

The Coulomb criterion applies to dry materials. For rock subjected to internal pore pressures, the outward pressure of the fluid reduces the effective normal stress on the solid matrix, and the rock fails at lower shear stress. The Mohr circles in Fig. 11.6(c) move to the left, corresponding to a reduction in normal stress. We replace normal stress in Eq. (11.36) with an effective normal stress σ_n^* given by the externally applied normal stress, σ_n, minus the internal stresses that are generated by pressurized fluids in the rock. The internal pore fluid pressure generates an outward directed stress, P_{F}, proportional to the pressure in the fluid and the area fraction of the fault plane occupied by pore space. In many sedimentary rocks, pore fluids are pervasive in wet grain–grain boundaries and P_{F} is nearly equal to the pore pressure. In more compact rock, with restricted fluid filled pores, P_{F} is generally smaller, although, in some cases, sealed pores may have high internal pressures. Equation 11.36 becomes

$$\tau = \mu_{\mathrm{f}}(\sigma_n - P_{\mathrm{F}}) + S_0, \tag{11.40}$$

written as

$$\tau = \mu_{\mathrm{f}}\sigma_n(1 - \lambda_{\mathrm{H}}) + S_0 \tag{11.41}$$

or

$$\tau = \mu_{\mathrm{f}}\sigma_n^* + S_0,$$

where λ_{H}, known as the Hubbert–Rubey coefficient, is the ratio of pressure from pore fluids to normal stress, and σ^* is the effective normal stress. Then

$$\lambda_{\mathrm{H}} = \frac{P_{\mathrm{F}}}{\sigma_n}. \tag{11.42}$$

Alternatively,

$$\tau = \mu_{\mathrm{f}}^*\sigma_n + S_0, \tag{11.43}$$

where $\mu_{\mathrm{f}}^* = \mu_{\mathrm{f}}(1 - \lambda_{\mathrm{H}})$ is the effective friction (Section 13.7).

We return to the Anderson criteria, which describe maximum stress directions for the different fault mechanisms. In a compressional regime where the maximum principal stress is horizontal, the dips of fault planes, relative to the surface, are expected to be shallow. In a normal faulting regime where the maximum stress is vertical, the dips of fault planes are expected to be steep. In a strike-slip regime faults should be vertical and angled closer to the direction of maximum principal stress than to the direction of minimum principal stress. However, geological and seismological observations of faults indicate that their orientations can be explained by these equations with a value of μ_{f} much smaller than suggested by laboratory measurements (0.6 to 0.85), as in Table 11.2. These criteria refer to mechanically isotropic material, but in many cases established faults are planes of weakness that control fault movements. Then the Anderson criteria refer to the direction of stresses that caused the faults in the first place and not necessarily to the present stress.

Table 11.2 Average fault dips for different friction coefficients compared with observations by Sibson and Xie (1998) and Jackson and White (1989)

	Dip ($\mu_{\mathrm{f}}=0.85$)	Dip ($\mu_{\mathrm{f}}=0.2$)	Observations
Normal fault	65.2°	50.7°	50.3°
Strike-slip fault	vertical, 24.8° to σ_1	vertical, 39.3° to σ_1	–
Reverse fault	24.8°	39.3°	39°

Sibson and Xie (1998) plotted a histogram of the dip angles of 31 reverse faults for which there are seismic or field estimates of the geometry of the fault planes (Fig. 11.8(a)). The mean of their broad distribution is 39°. Jackson and White (1989) present a histogram of dips for 15 normal fault events with a mean value of 50.3° (Fig. 11.8(b)). The modest number of events in each case invites doubt about the statistical significance of the small average deflections from 45°, but they are clearly incompatible with friction coefficients as high as the laboratory values of 0.6 to 0.85. While we know that established fault planes have powdered, soft material (gouge), compatible with a low friction coefficient, this

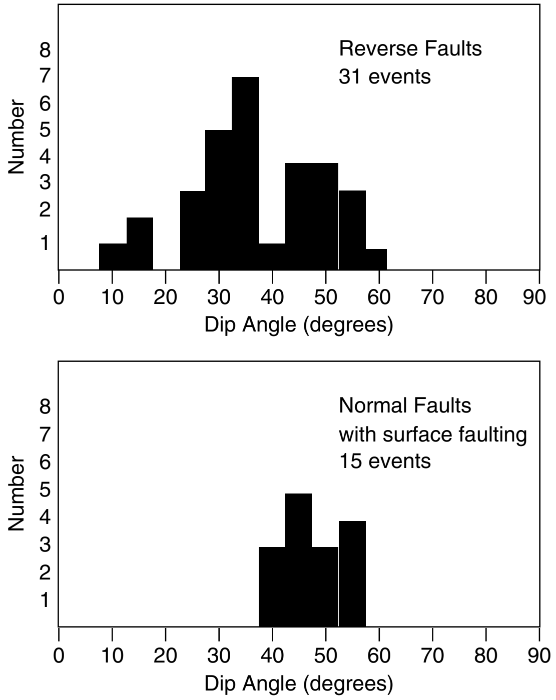


FIGURE 11.8 Histograms of measured dip angles of faults, as tests of the Anderson criteria (Table 11.2). (a) Reverse faults (mean dip angle 39°), after Sibson and Xie (1998). (b) Normal faults (mean dip angle 50.3°), after Jackson and White (1989).

does not explain how the faults formed in the first place and that is what matters in consideration of the angles. Evidently pore pressure in faults has a more significant effect that is yet to be satisfactorily explained. It is especially obvious that the conventional concept of friction can apply to only very shallow faults and needs modification to apply to earthquakes at even modest depths, where overburden pressure ensures very high normal stresses across faults.

11.6 Crustal stress: measurement and analysis

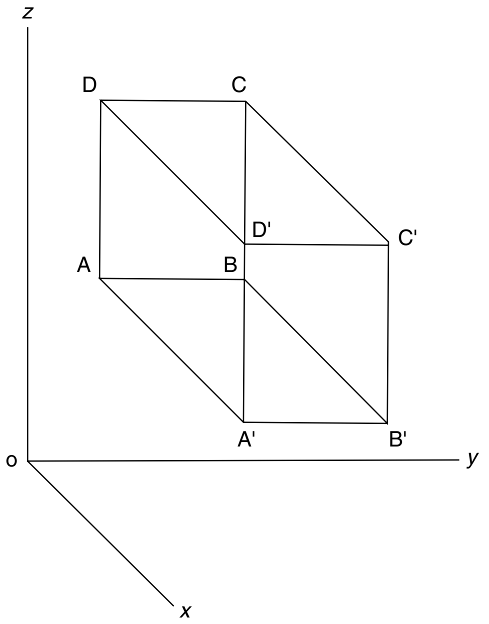


FIGURE 11.9 Geometry of an elementary volume in a stressed medium.

A basic starting point in generating a theoretical model of stresses is to find solutions that satisfy the equations of motion. The equations of motion in a continuous medium are found by applying Newton's third law to elementary volumes subjected to a stress field σ_{ij}. Consider a parallelepiped of length δx and cross-sectional area $\delta y\,\delta z$ (Fig. 11.9). The normal force acting on the face ABCD is

$$F_1 = \sigma_{xx}\delta y \delta z, \tag{11.44}$$

and on face A′ B′ C′ D′,

$$F_2 = \left\{\sigma_{xx} + \frac{\partial \sigma_{xx}}{\partial x}\delta x\right\}\delta y \delta z. \tag{11.45}$$

The net force from this stress component is:

$$F = F_2 - F_1 = \frac{\partial \sigma_{xx}}{\partial x}\delta x \delta y \delta z. \tag{11.46}$$

If we consider the shear stresses on the other two sides in a similar fashion we obtain additional forces in the x-direction of

$$\frac{\partial \sigma_{yx}}{\partial y}\delta x \delta y \delta z, \quad \frac{\partial \sigma_{zx}}{\partial z}\delta x \delta y \delta z.$$

Summing these forces, the net force per unit volume in the x-direction due to the stress imposed on the parallelepiped by the material exterior to it is

$$\frac{\text{force}}{\text{volume}} = \frac{\partial \sigma_{xx}}{\partial x} + \frac{\partial \sigma_{xy}}{\partial y} + \frac{\partial \sigma_{xz}}{\partial z}. \tag{11.47}$$

In addition to these external forces, the volume may be subjected to an internal body force, most importantly gravity, mg, which must be added to the above equation. Let X be the component of the body force per unit mass (g in the gravity case). Then the force per unit volume is ρX. The equation of motion becomes force/volume = density × acceleration. For three dimensions

$$\begin{aligned}
&\frac{\partial \sigma_{xx}}{\partial x}+\frac{\partial \sigma_{yx}}{\partial y}+\frac{\partial \sigma_{zx}}{\partial z}+\rho X=\rho \ddot{u}_x,\\
&\frac{\partial \sigma_{xy}}{\partial x}+\frac{\partial \sigma_{yy}}{\partial y}+\frac{\partial \sigma_{zy}}{\partial z}+\rho Y=\rho \ddot{u}_y,\\
&\frac{\partial \sigma_{xz}}{\partial x}+\frac{\partial \sigma_{yz}}{\partial y}+\frac{\partial \sigma_{zz}}{\partial z}+\rho Z=\rho \ddot{u}_z.
\end{aligned} \tag{11.48}$$

Solutions to these differential equations are used in numerous applications from seismology to structural geology. In this section we consider the equilibrium versions with no acceleration.

The equations of motion (11.48) may also be expressed in cylindrical or spherical coordinates (Jaeger and Cook, 1984). The equilibrium equations for horizontal plane stress in cylindrical coordinates are

$$\begin{aligned}
&\frac{\partial \sigma_{rr}}{\partial r}+\frac{1}{r}\frac{\partial \sigma_{\theta r}}{\partial \theta}+\frac{\sigma_{rr}-\sigma_{\theta\theta}}{r}=0,\\
&\frac{\partial \sigma_{r\theta}}{\partial r}+\frac{1}{r}\frac{\partial \sigma_{\theta\theta}}{\partial \theta}+\frac{2\sigma_{r\theta}}{r}=0,
\end{aligned} \tag{11.49}$$

where the body forces are omitted because we are considering (r,θ) geometry in the horizontal plane perpendicular to the gravitational force.

Consider a cylindrical hole of radius R in an elastic medium subjected to a pressure P. The solution that satisfies Eq. (11.49) is

$$\begin{aligned}
\sigma_{rr}&=P\frac{R^2}{r^2},\\
\sigma_{\theta\theta}&=-P\frac{R^2}{r^2},\\
\sigma_{r\theta}&=0,\ \text{for } r>R.
\end{aligned} \tag{11.50}$$

This solution describes stresses around pressurized conduits such as a magma conduit or a pressurized borehole.

A related solution, and one of the most important in rock mechanics, gives the stresses outside a pressurized cylindrical hole with specified principal stresses $\{\sigma_1,\ \sigma_3\}$ at infinity (Jaeger and Cook, 1984). It is the theoretical basis for interpreting stress measurements in the Earth using hydrofracture, overcoring and borehole breakouts.

$$\begin{aligned}
\sigma_{rr}&=P\frac{R^2}{r^2}+\frac{1}{2}(\sigma_1+\sigma_3)\left(1-\frac{R^2}{r^2}\right)\\
&\quad+\frac{1}{2}(\sigma_1-\sigma_3)\left(1-\frac{4R^2}{r^2}+\frac{3R^4}{r^4}\right)\cos(2\theta),\\
\sigma_{\theta\theta}&=-P\frac{R^2}{r^2}+\frac{1}{2}(\sigma_1+\sigma_3)\left(1+\frac{R^2}{r^2}\right)\\
&\quad-\frac{1}{2}(\sigma_1-\sigma_3)\left(1+\frac{3R^4}{r^4}\right)\cos(2\theta),\\
\sigma_{r\theta}&=-\frac{1}{2}(\sigma_1-\sigma_3)\left(1+\frac{2R^2}{r^2}-\frac{3R^4}{r^4}\right)\sin(2\theta),
\end{aligned} \tag{11.51}$$

where θ is measured from the direction of σ_1.

Hydrofracture is the process whereby water pressure in a conduit, such as a borehole, becomes large enough to fracture the surrounding rock, forming a crack into which the water flows. Hydrofracture occurs in natural systems such as in geothermal regions. The equivalent process, magma-fracture, occurs in volcanic areas where magma fractures rock to generate dikes and sills. Hydrofracture is used in the oil industry to fracture oil reservoirs, and to increase permeability for secondary recovery of reserves. The hydrofracture process involves sealing off a segment of the borehole with packers, and pumping liquid into the sealed section until the surrounding rock breaks. The pressure is monitored throughout, and various methods are used to estimate the directions and orientations of fractures. In addition to its oil reservoir use, the method is one of the most important in determining the regional stress field, because fractures open in the direction of minimum principal stress, and fluid flow into the cracks is governed by their dimensions.

In order to initiate a vertical fracture, for example, the tangential stress, $\sigma_{\theta\theta}$, at the borehole, $r=R$ in Eq. (11.51), must be equal to the tensional strength of the rock, $-T_0$,

$$\sigma_{\theta\theta}=-T_0.$$

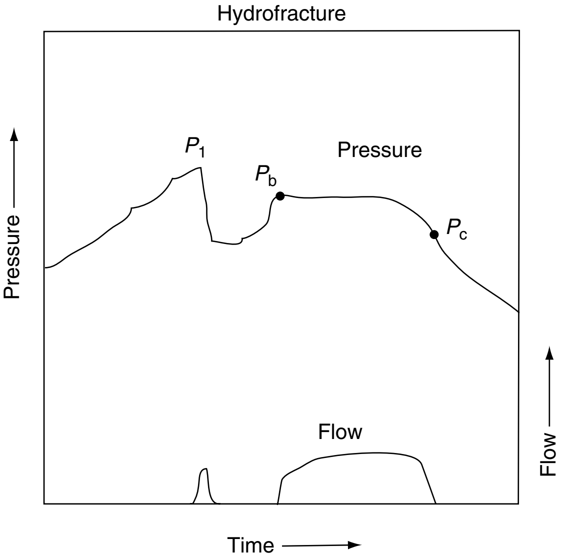


FIGURE 11.10 Pressure at the well-head for a hydrofracture experiment. Pressure in the borehole is increased to P_1 until the rock fractures, fluid flows into the crack and the pressure decreases. When the flow stops the crack closes and pressure increases again to P_b, which is the breakout pressure at which the crack re-opens to flow. P_c is the shutin or crack closure pressure when flow stops.

From Eq. (11.51) we obtain at $r = R$,

$$\sigma_{\theta\theta} = -P + (\sigma_1 + \sigma_3) - 2(\sigma_1 - \sigma_3)\cos(2\theta).$$

The tangential stress becomes most tensile (negative) at $\theta = 0$. Then

$$T_0 = P + \sigma_1 - 3\sigma_3, \tag{11.52}$$

i.e. a crack forms at $P_1 = T_0 + 3\sigma_3 - \sigma_1$ (Fig. 11.10) and water flows from the borehole into the crack. Before the crack becomes very large and influences the regional stresses by its presence, the pressure is reduced, flow ceases, and then the pressure is applied once more. This time the internal pressure inflates an already broken crack of no tensile strength and we obtain the simpler equation

$$P_b = 3\sigma_3 - \sigma_1. \tag{11.53}$$

P_b is the breakdown pressure, which is measured at the well-head. Flow resumes at $P = P_b$ (Fig 11.10). We thus have one equation, Eq. (11.53), in two unknowns, σ_1 and σ_3, and require a further measurement. The crack is allowed to grow. If the material is impermeable and isotropic, the usual assumption, the crack will open normal to σ_3 and extend in the direction of σ_1. After the crack has been fully inflated, the pressure at the well-head is reduced, and the crack collapses, reducing the flow to zero. For a large crack σ_1 has no influence on the flow, and the internal pressure is dependent on $P - \sigma_3$. The crack closes when $P - \sigma_3 = 0$. Thus the shutin or closure pressure, P_c, measured just before closure, gives an estimate of σ_3. In many situations the third principal stress can be taken as $\sigma_2 = \rho g z$. Then the hydrofracture measurement can, in principle, determine all the components of the stress field. The direction the crack travels is measured by in-borehole methods involving televiewers, impression packers, or externally by detecting the deformation field at the surface by arrays of very sensitive tiltmeters (Davis, 1983).

Another method of estimating stresses, known as overcoring, involves drilling out the rock in a cylindrical annulus around a borehole, effectively isolating it from the tractions of the surroundings. Calipers are used to measure the change in shape as a function of depth after overcoring. The associated strains may be converted to stresses if the elastic constants have been determined so that Eq. (11.51) can be applied to determine σ_1 and σ_3.

Also, Eq. (11.51) is applied to the analysis of borehole breakouts. These are localized failure of the borehole walls in which rock spalls off along sections of the circumference closest to the axis of minimum principal stress σ_3. A squeezed hole breaks along its sides. Equation (11.51) with $P = 0$ describes the stress concentration that occurs when a hole is drilled into a region of principal stresses σ_1, v_1, v_3, σ_3. The greatest compression in the rock occurs for $\theta = \pm 90^\circ$, which for $r = R$ gives $\sigma_{\theta\theta} = 3\sigma_1 - \sigma_3$. The compressive stress causes cracks that are parallel to the circumference and flakes of rock spall from each side leaving the bore elongated in the horizontal plane. The direction of minimum principal stress is mapped by inserting a caliper and measuring the orientation of the elongation. While useful in giving stress directions, this method gives no information on the magnitudes of the stresses.

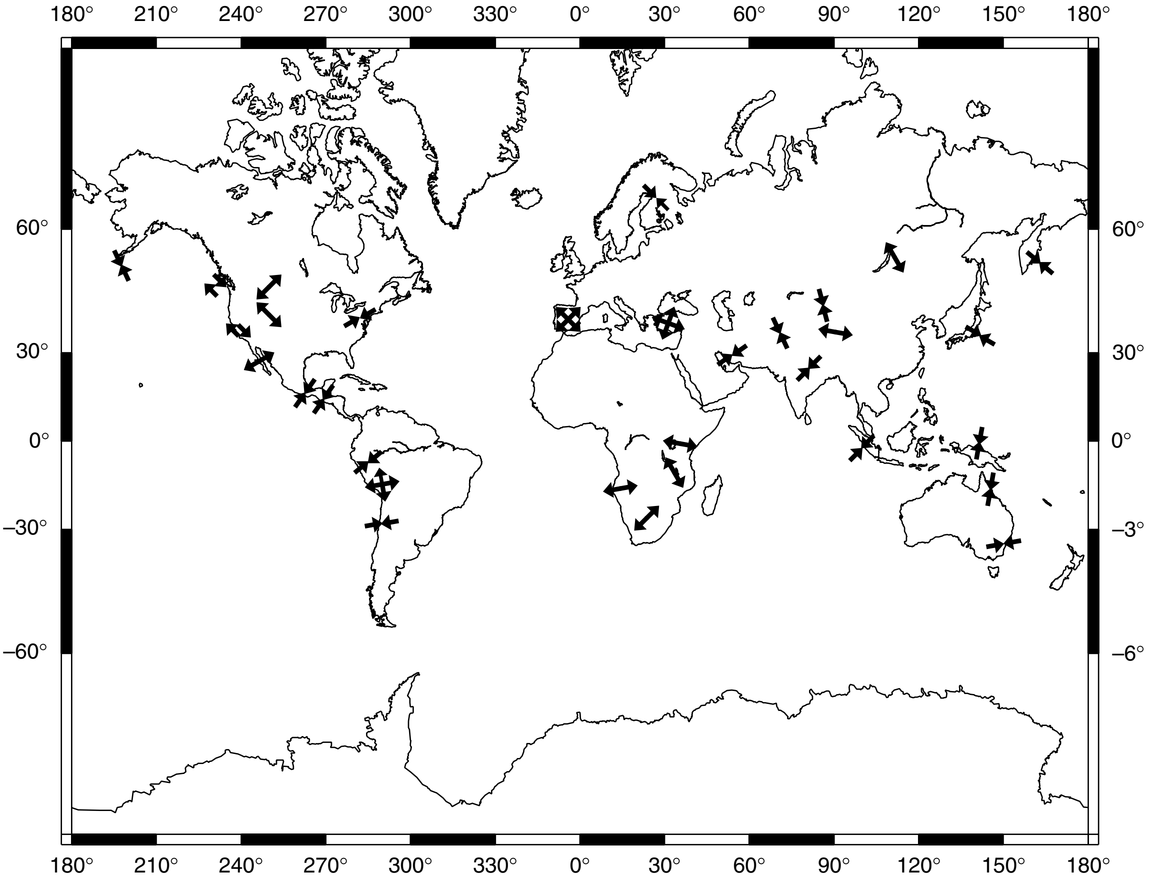


FIGURE 11.11 Selected details from the world stress map by Reinecker *et al.* (2004). Diverging arrows represent regions of extension and converging arrows indicate compression. Anti-parallel pairs of arrows denote strike-slip faulting.

Figure 11.11 is a plot of stress regimes from the world stress map (Reinecker *et al*, 2004), which combines hydrofracture, overcoring and borehole breakout measurements with earthquake focal mechanisms (Chapter 14) and geological indicators such as faulting type. Large regions of the world exhibit coherent stress directions. Diverging and converging arrows represent regions of tension or compression, respectively. Adjacent oppositely directed arrows indicate strike-slip regions. For example, the Aleutian subduction zone is a thrust regime, whereas the San Andreas fault is strike-slip, but further inland, the Basin and Range province of west North America is a normal faulting regime, as is Baja California, where crustal extension is taking place. The major active continental rift zones of the world, which include the East African rift in Kenya, the Rio Grande rift in New Mexico, and Lake Baikal, Siberia, are all in regions of extensional stress. The spatial coherence of stress can be explained by a combination of large-scale stresses from plate tectonics and buoyancy stresses associated with uplift, as considered further in Sections 13.4 and 13.6.

12

Tectonics

12.1 Preamble

Seismicity (seismic activity) is a word coined by Gutenberg and Richter (1941) to encompass earthquake occurrences, their mechanisms, magnitudes and especially their geographical distribution. Although we have known for more than 150 years that earthquakes are concentrated in extended, but relatively narrow bands across the Earth (Fig. 12.1), the pattern remained more or less mysterious until it became a cornerstone of the theory of plate tectonics in the 1950s and 1960s. According to this theory the surface of the Earth is divided into almost rigid plates that are in relative motion, with earthquakes occurring mainly at the boundaries. Especially significant in this connection are the deep focus earthquakes, which mark the subduction zones where cooled surface material plunges into the mantle. They provide direct evidence of the deep, convective motion that drives the plates.

There are now more than 3000 globally distributed seismological stations routinely contributing data to the International Seismological Centre at Thatcham, UK. Although they are unevenly distributed, they suffice to locate reliably all earthquakes of magnitude 5 or greater (the definition of magnitude and its relationship to energy are discussed in Section 14.6). By restricting attention to these events we can view the pattern of world seismicity without a bias towards instrumented areas. Earthquake epicentres (Fig. 12.1), that is the surface points directly above the foci (hypocentres) where earthquakes actually occur, outline the plates, which are identified in Fig. 12.2. Intraplate earthquakes also occur, although much less frequently, demonstrating that the plates are not completely rigid. However, plate deformation is slight enough to neglect in calculations of their relative motions. The pattern of mantle convection suggested by the distribution of earthquakes in Fig. 12.1 is confirmed by studies of the first motions of seismic waves (Section 14.4), which are used to infer the relative motions of the plates at their common boundaries.

The development of our understanding of global scale tectonics was pioneered by paleomagnetism, which established the credibility of continental drift (Section 25.6). Subsequently, laser ranging to satellites (SLR), very long baseline interferometry (VLBI), and, now much more extensively, the global positioning system of satellites (GPS) have been used to show that the geologically inferred relative motions of continental blocks are on-going. Another crucial contribution by paleomagnetism was the establishment of the sequence of geomagnetic reversals (Section 25.4). Fresh igneous crust produced at the spreading ocean ridges is magnetized during or shortly after its appearance and carries a record of the field with it as it moves away from the ridges. This has produced a series of linear ocean floor magnetic anomalies parallel to the ridges, correlated with the irregular alternations of the polarity of the geomagnetic field seen in the magnetism of continental rocks. For the last few million years the reversals have been dated isotopically with sufficient accuracy to determine the rates of ocean floor spreading from the

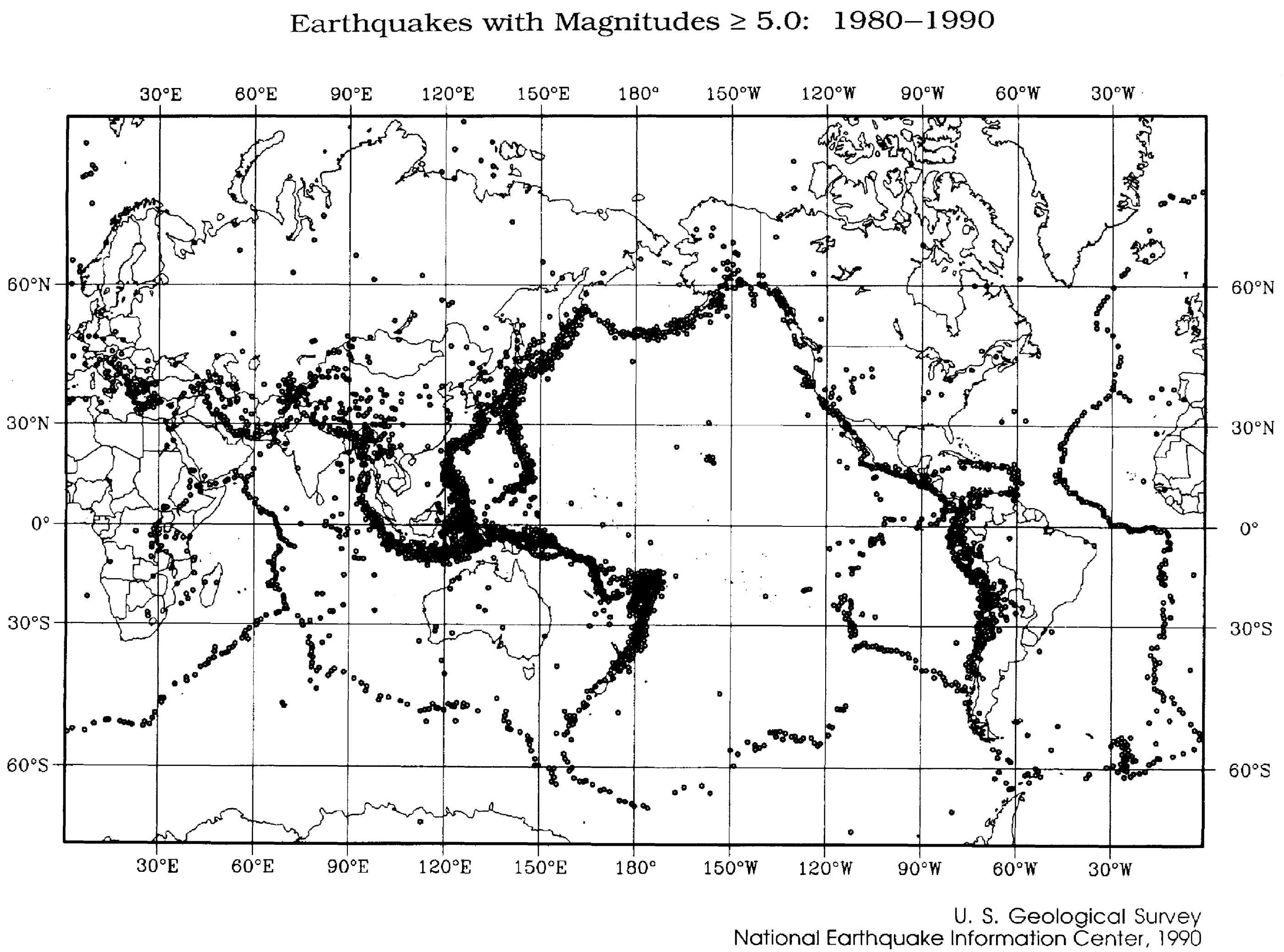


FIGURE 12.1 Epicentres of earthquakes with magnitudes exceeding 5.0 that occurred in 1980–1990. National Earthquake Information Center, US Geological Survey, Denver, courtesy of Susan K. Goter.

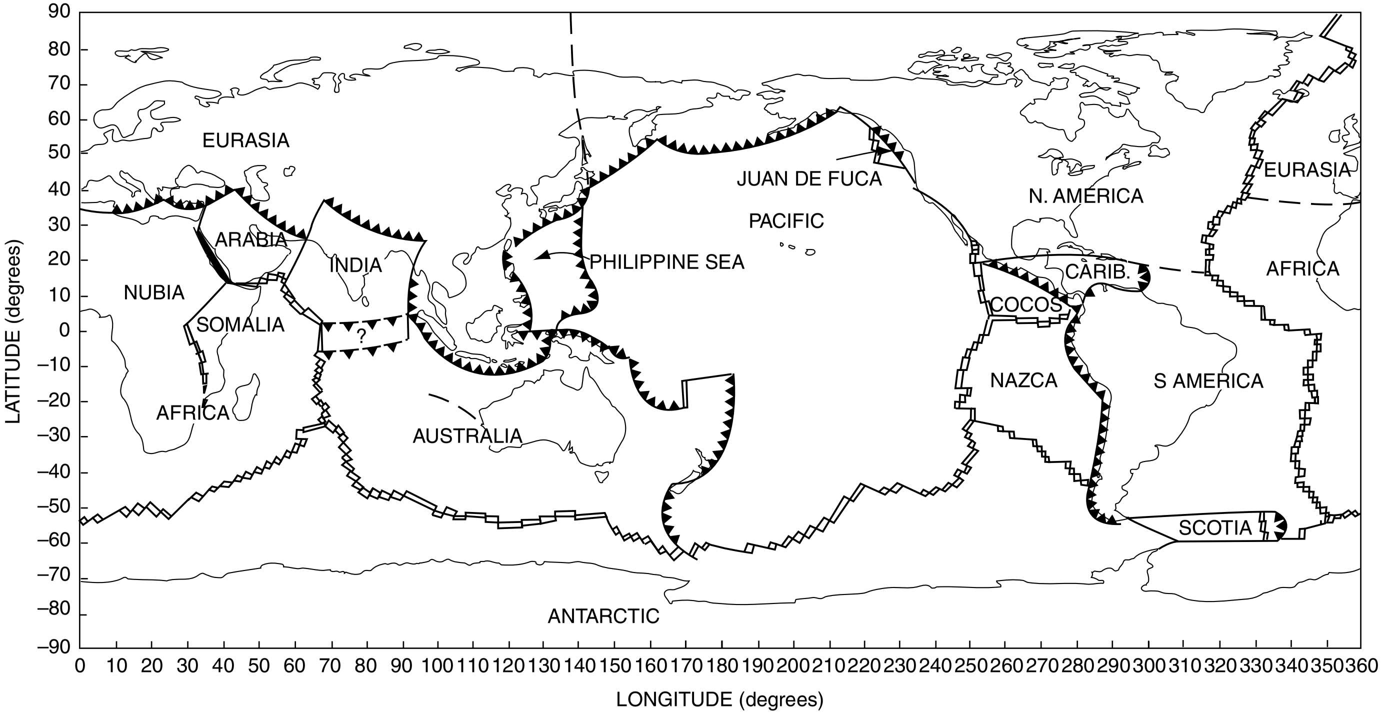


FIGURE 12.2 The Earth's major plates. Subduction zones are represented by 'shark's teeth' drawn on the over-riding plates and showing the motion of the subducting plates beneath. Spreading centres are marked by double lines, but they are fragmented by transform faults (single lines, as in Fig. 12.7). Broken lines mark uncertain boundaries.

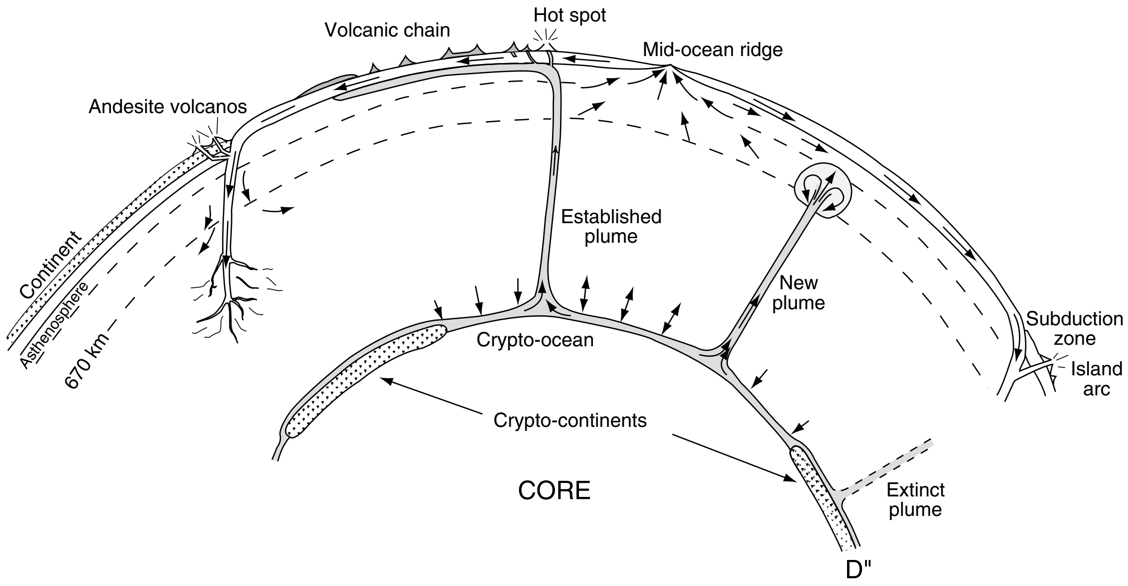


FIGURE 12.3 A pictorial cross-section of mantle convection.

spacing of the magnetic stripes. For the period back to 100 million years or so the magnetic polarity time scale is established by the sea floor anomalies, with the assumption that the spreading rate is constant.

A third paleomagnetic observation with fundamental implications for tectonics is the motion of the plates across volcanic 'hot spots' that are anchored deep in the mantle. The idea of a mantle reference frame originated with observations made in Hawaii. The presently active volcanos are on Hawaii Island at the south-east end of the island chain, latitude 19°, and at Loihi, a submarine volcano immediately to the SE of that. The other islands in the chain follow a line in the north-west direction and have ages that increase linearly with distance in that direction. Paleomagnetic measurements showed that they were all formed at latitude 19°, that is, at the latitude of the present centre of volcanism. The Pacific plate is moving across a more or less stationary mantle hot spot that produces a series of volcanos, with their extinct and eroding edifices steadily drifting north-west.

The concept of deep mantle convective plumes, with local hot spots, such as Hawaii, as their surface expressions, originated with Morgan (1971) and has become an important component of our understanding of mantle convection, as illustrated in Fig. 12.3. It is not certain how many there are because they are not all equally active. The four most active occur in quite different situations. Hawaii and Reunion are oceanic islands, Iceland sits astride the mid-Atlantic ridge and a mid-continent hot spot occurs in Zaire, central Africa. Yellowstone is a commonly cited example, but appears to be physically different from the others and may have a somewhat different cause. Surface traces of several hot spots are recognized, including some others on the Pacific plate in lines roughly parallel to the Hawaiian chain. Those on other plates have independent traces oriented according to their plate motions. The apparent fixity of the Hawaii hot spot invited the supposition that they are all more or less stationary with respect to an immobile deep mantle and provide a reference frame for plate tectonics. It is now clear that they are moving with respect to one another, but at speeds lower than the plate motions and that they give an indication of convective motion deep in the mantle. Particularly relevant is a check on the relative motion of the two major oceanic hot spots that are remote from plate boundaries, Hawaii and Reunion (see Section 12.4).

The discoveries of paleomagnetism made plate tectonics a convincingly quantitative explanation of global geological processes, but many of the ideas have long histories. The concept of mantle convection can be traced back to the early nineteenth century when it was supposed that the deep interior is largely fluid. Even with the mantle recognized to be solid, sea floor spreading and subduction were advocated by A. Holmes and a figure showing these processes in a form very similar to our present understanding appeared in the 1944 edition of his textbook *Principles of Physical Geology* and is reproduced in Cox (1973, p. 20). These ideas were brought into focus by echo-sounding across the Pacific, especially by H. Hess, a submarine commander during World War II, who subsequently used his observations to link continental drift with sea floor spreading. Plate tectonics was generally accepted only when the structure of the ocean floor had become clear.

Although comprehension of global geology was very incomplete before it incorporated the ocean floors, it is also true that their restricted age range would present a very limited view of Earth history without evidence from the continents. The permanence of the continents is due to their buoyancy, which prevents subduction. The continental crust has an average thickness $z = 37\,\mathrm{km}$ and an average density contrast relative to the underlying mantle $\Delta\rho \approx 500\,\mathrm{kg\,m^{-3}}$, giving a mass deficiency (relative to an equal volume of mantle material) $z\Delta\rho \approx 1.85 \times 10^7\,\mathrm{kg\,m^2}$. This is greater than the negative buoyancy of thermally contracted lithosphere, which has shrunk by $\Delta z \approx 2.1\,\mathrm{km}$; taking the mantle density to be $\rho = 3350\,\mathrm{kg\,m^{-3}}$, the mass excess is $\rho\Delta z = 7.0 \times 10^6\,\mathrm{kg\,m^2}$. The negative buoyancy of cooled lithosphere fails by a factor exceeding two to overcome the positive buoyancy of continental crust. Only oceanic sections of plates subduct and when they bring continental blocks together, collision zones appear, most impressively the Himalaya. Even for oceanic plates there is some crustal buoyancy, but it is not clear that it seriously inhibits subduction of young oceanic lithosphere. At depth, basalt converts to eclogite and loses its buoyancy, so that the resistance to subduction is probably a shallow effect.

12.2 Wadati–Benioff zones and subduction

Deep earthquakes were first clearly identified in 1928 from Japanese records studied by K. Wadati. They are now recognized to occur down to 700 km along the zones of strong subduction, shown in Fig. 12.2, and to mark inclined planes extending into the mantle from the zones of convergence of the lithospheric plates. Deep ocean trenches mark the lines where the subducting plates turn down. The contribution of H. Benioff to the identification of planes of deep seismicity is acknowledged by referring to them as Wadati–Benioff zones. By convention earthquakes are classified into three groups, with shallow (0 to 60 km), intermediate (60 to 300 km) and deep (>300 km) foci, although it is sometimes convenient simply to refer to all earthquakes with foci below 60 km as deep. Shallow earthquakes are the most numerous. The largest earthquakes occur at shallow depths in subduction zones.

The detailed geometries of the Wadati–Benioff zones are variable and depend on factors such as plate speed, age of the subducting lithosphere and its geology, especially the distribution of oceanic and continental crust. The deep seismicity of Japan has been studied in particularly close detail and gives valuable insight on the subduction process. Figure 12.4 shows the distribution of deep earthquakes in this area, where there are intersections of several subduction arcs. Figure 12.5 shows an east–west cross-section of a subset of the data in Fig 12.4, between latitudes 39°N and 40°N. In this case, precise location of foci by a close network of seismic stations has delineated a pair of parallel planes. Focal mechanisms for earthquakes show that the upper plane is in down-dip compression and the lower one is in down-dip extension. These observations are naturally explained by identifying the upper plane as the upper boundary of the subducting lithospheric slab, and the lower plane as the middle of the slab, which fails in tension due to the negative buoyancy that is carrying the plate down. In other subduction zones, down-dip extension dominates

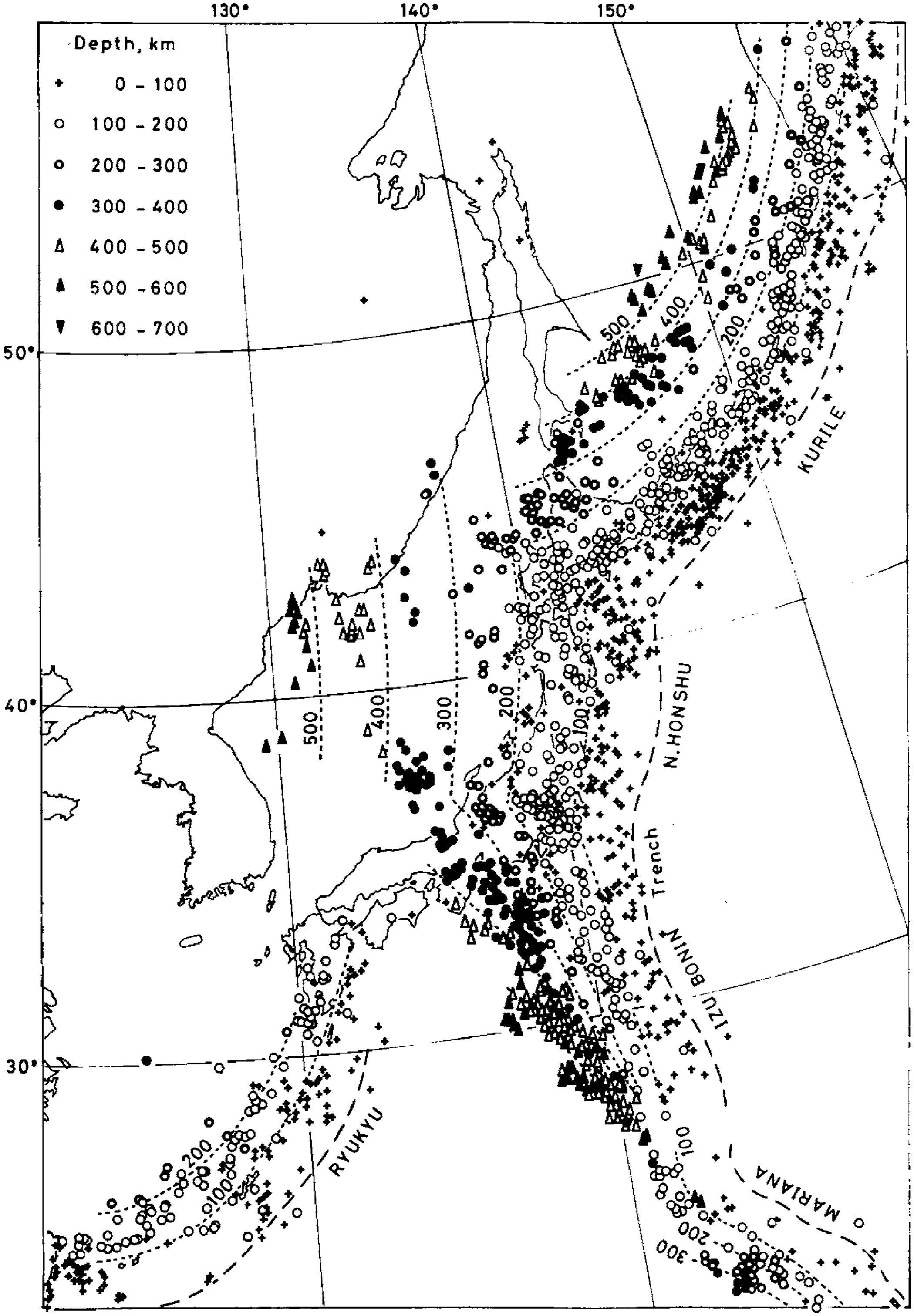


FIGURE 12.4 The pattern of earthquake foci down the Wadati–Benioff zones of Japan and neighbouring arcs. Reproduced, by permission, from a drawing by Sasatani (1989) of an original plot by T. Utsu. Several inclined planes of foci intersect in this area, indicating complicated plate geometry.

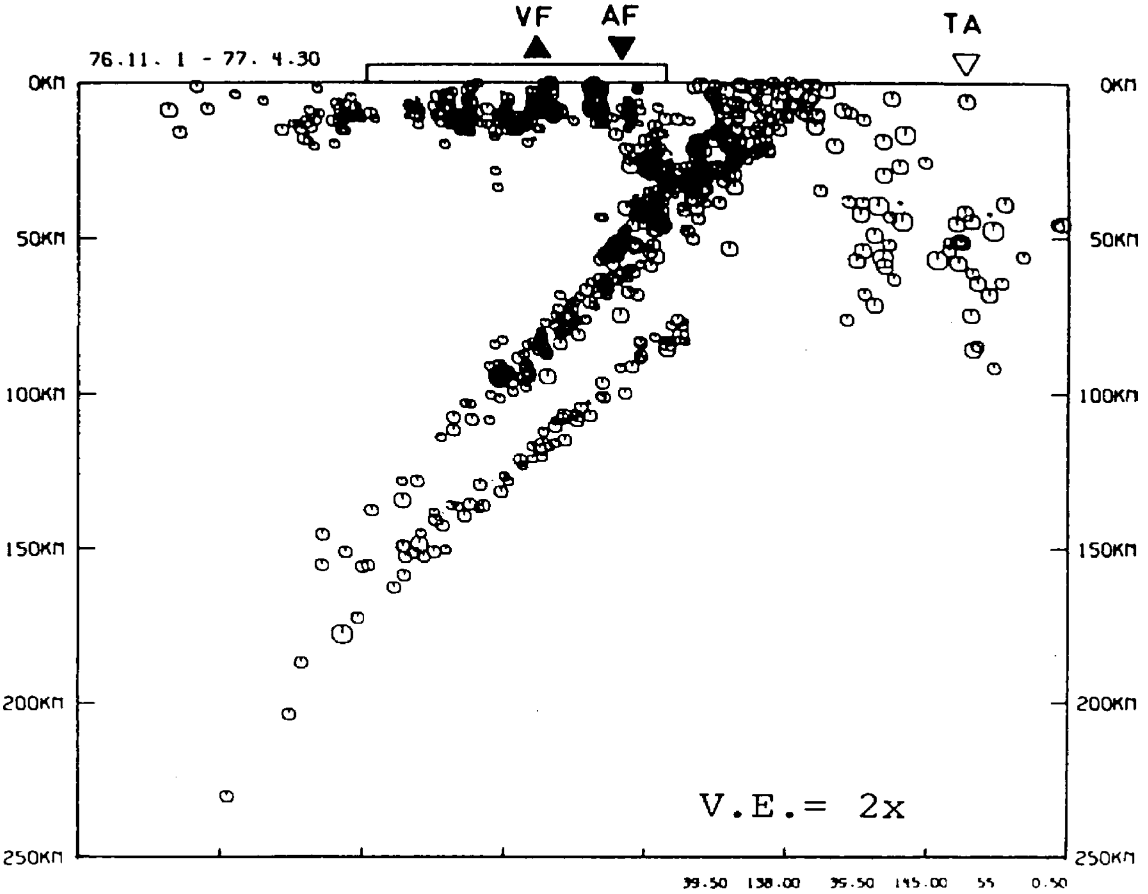


FIGURE 12.5 Cross-section of the Wadati–Benioff zone under northern Honshu, Japan, showing two parallel planes of earthquake foci. VF indicates the volcanic front, at the centre of the land area. Reproduced, by permission, from Hasegawa (1989). For a colour plot with greater detail see Hasegawa *et al.* (1991).

at intermediate depths, but the deepest shocks give down-dip compression, indicating increasing resistance by mantle viscosity.

An intriguing feature of subduction zones is the wide range of angles of subduction. At one extreme it appears to be almost horizontal for some distance, for example under central Peru (Fig. 12.6). Steep subduction is slightly more common, with an average angle of about 50°. There have been several attempts to explain this variability, and the fact that the subducting slabs are not vertical, although they are driven by gravity. None of the explanations is yet convincing.

There is considerable variability in the seismic activity in subducting material, due to differences in local conditions that appear to be reflected in volcanic activity. Deep earthquakes are much rarer than shallow ones everywhere, and in some places they do not occur at all. There is a similar variability in the volcanic activity along subducting arcs. As indicated in the cartoon of tectonic processes in Fig. 12.3, subducted material generates andesitic lava (named after the Andes Mountains in South America), producing lines of volcanos which appear where the slabs have penetrated to about 100 km depth. But along some sections of the subduction zones volcanos are completely absent, including central Peru (Fig. 12.6). Either the chemistry or the mode of subduction is unsuitable for magma generation in these areas. It may be that one important factor is the availability of water in subducted sediments. The section of the Nazca plate that is subducting under central

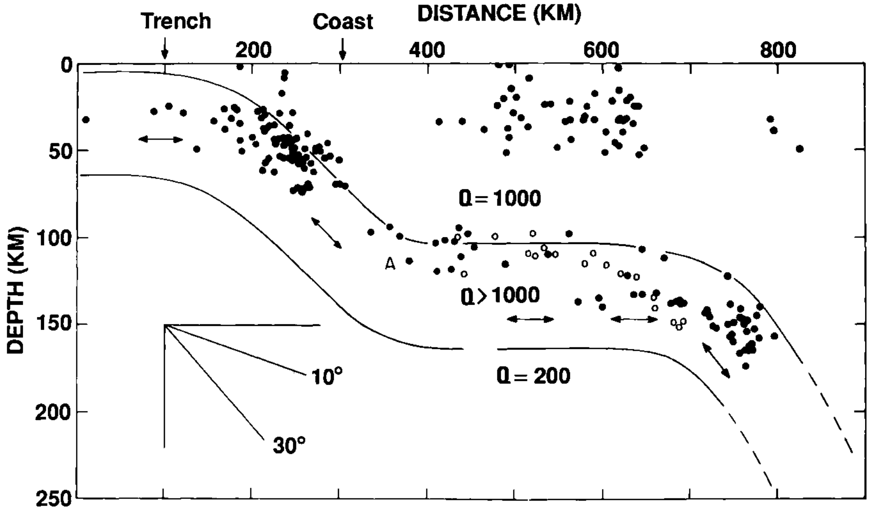


FIGURE 12.6 A cross section of the Wadati–Benioff zone under central Peru, showing evidence of 300 km of 'horizontal subduction'. Solid circles are earthquake foci plotted by M. Barazangi and B. Isacks for a 100 km wide section. Schneider and Sacks (1992) added the open circles, representing foci determined by a local network for a more restricted area. Arrows indicate extensional stress deduced from focal mechanisms.

Peru includes an atypical crustal component, that may be a continental fragment, the Nazca ridge. Being more elevated than most of the ocean floor, it has accumulated less wet sediment. Another possibility is that, if subduction is slow or delayed, as in the central Peru case, the volcanic heat is too diffuse for magma generation and slower, broader scale upwelling occurs, with granitic intrusions.

The incorporation of subducted material in the lava produced by subduction zone volcanos was proved beyond all doubt by the discovery of the isotope ^{10}Be in andesitic lavas (see especially Morris *et al.*, 1990). This is a radioactive isotope produced in the atmosphere by cosmic ray bombardment, washed into the sea and deposited in marine sediments. Its half-life, 1.5×10^6 years (Table H.2, Appendix H), is short enough to ensure that it could not have a long residence time in the Earth. The lava incorporates marine sediment that cannot be more than 10^7 years old. Subduction, to 100 km at least, must proceed at the full speed of the surface plates and, in some areas, all the sediment must be subducted.

Morris *et al.* (1990) pointed out also that the abundance of boron in andesitic lavas requires the incorporation of sea water as an important component. There is no other plausible source of boron. Thus we arrive at the conclusion that andesitic volcanism is a consequence of the subduction of sea water. Melting of dry material would not occur and it is the availability of water that leads to magma generation. Since the Earth's free ocean water is unique in the Solar System, it follows that its acid volcanism is also unique, at least at the present time. As far as we know, all lavas on the other terrestrial planets (and the Moon) are basaltic. In Sections 1.1 and 2.9 it is pointed out that acid rocks are the essential ingredients of continents, which remain as rafts on the surface as the mantle convects. The subduction of sea water is responsible for the development of continental crust and therefore for the bimodal surface elevation shown in Fig. 9.4.

The ultimate fate of subducted material has been a subject of conjecture. Since earthquakes do not occur below 700 km the evidence for the penetration of slabs beyond that is less direct. This is approximately the depth of a major phase transition, accepted as the boundary between the upper and lower mantles. Viscosity increases with depth and the lower mantle is almost certainly more viscous than the upper mantle, but the thermal argument in Section 22.4 requires convection to be a whole-mantle process. Much of the Earth's heat originates in the lower mantle so the coolness carried downwards by subducted slabs must be distributed through the lower mantle. The problem of the mechanism is addressed in Section 12.6. Although the 660 km transition presents an obstacle to convection, in the absence of a strong mid-mantle thermal boundary whole mantle convection must prevail, perhaps with a delay or complication at 660 km, and this is emphasized by a consideration of lower mantle cooling (Section 12.6). In the upper mantle there can be little doubt that subducting slabs are continuous; observed breaks in seismic activity must be interpreted in terms of material properties and not discontinuities in the slabs themselves (Okal, 2001).

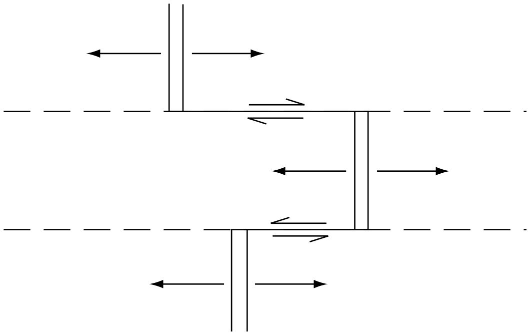


FIGURE 12.7 Transform faults. These are planes of horizontal shear between sections of a spreading centre, marked by double lines. The ocean ridges, where spreading occurs, are broken up by transform faults, which leave their signatures on the ocean floors as seismically inactive 'fracture zones', indicated by broken lines.

12.3 Spreading centres and magnetic lineations

The buoyancy of the continents is indicated by their 5 km elevation above the ocean floors (Fig. 9.4) and, as explained in Section 12.2, they resist subduction, so that where converging plates bring continents into collision, as at the Himalayan boundary between India and the main Asian mass, they form a crumpled pile of abnormally thick continental crust. Continents break up and reform in different ways, but they do not disappear. They gradually increase in total volume as acid volcanism adds to them, with the loss of minor fragments too small to escape subduction and sediments washed into the sea and subducted from there. Continents just grow older and they have cores of ages 3.5×10^9 years or more, but there is a continuous recycling via erosion and sedimentation, as discussed in Section 5.3.

The ocean floors are, by comparison, youthful. Nowhere is there ocean floor of age exceeding about 2×10^8 years and the average age at subduction is about 9×10^7 years. Since the ocean floors occupy 60% of the Earth's surface (3×10^8 km^2), their limited age-range is a measure of the rate at which oceanic crust is both forming and disappearing: $3 \times 10^8\,\text{km}^2/9 \times 10^7$ years $\approx 3.4\,\text{km}^2$/year. The ocean floors are a thermal boundary layer, transferring to the ocean the heat that is convectively transported from the deep mantle.

The ocean ridges, which mark the crustal spreading centres, are less active seismically than the subduction zones, but still show clearly in Fig. 12.1. However, they have no deep earthquakes and many of the shallow ones occur on transform faults, offsets between sections of ridge, rather than on the ridge axes where the spreading occurs (Fig. 12.7). In Section 13.2 it is argued that the spreading is a more-or-less passive consequence of subduction-driven convention and that mantle viscosity estimates can be rationalized only by accepting that plate motion is accommodated by shear in a relatively thin low-viscosity layer (the asthenosphere) and not by bulk mantle motions. Thus, most of the fresh crustal material appearing at the spreading centres probably flows from the adjacent asthenosphere. It is unlikely to be derived from more than modest depths, except for a possible asthenospheric connection between spreading centres and the deep mantle plumes that bring up core heat. Beneath spreading centres there is, therefore, nothing comparable to the Wadati–Benioff zones. Moreover, since there is no large, cool mass at a spreading centre, there is rather little material in which earthquakes are readily generated. This means that we do not have such a clear or detailed picture of the geometry of spreading as we do of subduction.

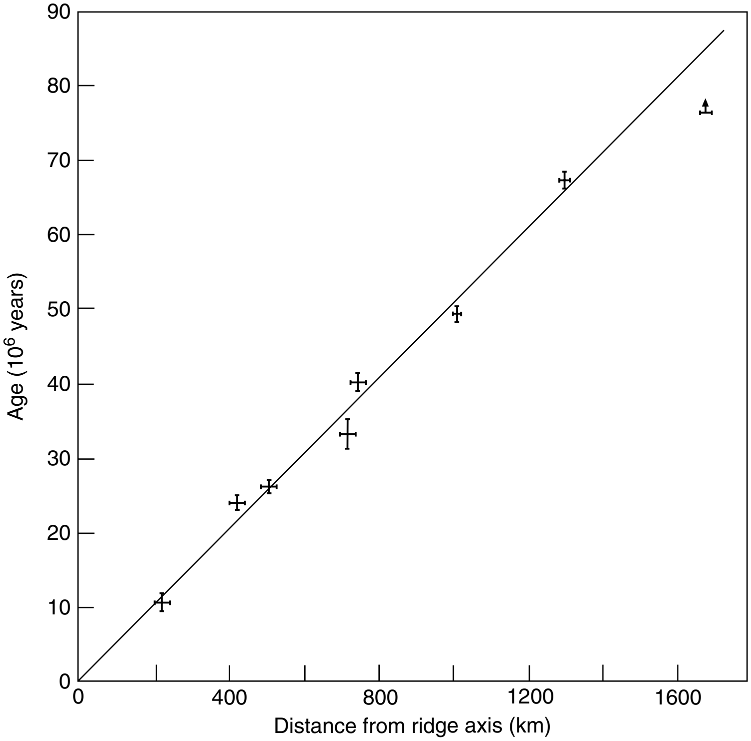


FIGURE 12.8 Paleontologically determined ages of the lowest ocean floor sediments, immediately above the basalt basement, which they date, as a function of distance from the mid-Atlantic ridge at 30 °S. Replotted from data by Maxwell *et al*. (1970).

Figure 12.8 shows the increase with distance from the centre of the mid-Atlantic ridge of the age of the ocean floor in the South Atlantic. Originally plotted by Maxwell *et al.* (1970), these data are probably still the most convincing evidence that the mid-Atlantic ridge is a spreading centre. The data points are mostly to the west of the ridge, but span it and demonstrate that the spreading is symmetrical, at 1.9 cm/year each way. More comprehensive evidence of the spreading speeds along all of the active ridges, shown in Fig. 12.2, has been obtained by comparing the spacing of the sequences of linear magnetic anomalies that they have produced. The ridge axes are generally marked by rifts, along which fresh lava appears, as the sides of rifts slowly move apart, with a 'tape recording' of the field polarity (Fig. 12.9). The linear magnetic anomalies were first recognized in the Pacific by A. D. Raff and R. G. Mason and their interpretation was given by Vine and Matthews (1963). The symmetrical pattern of anomalies across a part of the mid-Atlantic ridge is reproduced in Fig. 12.10.

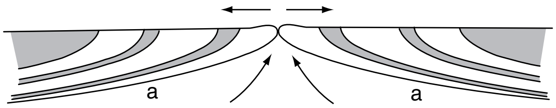


FIGURE 12.9 Alternating normal and reversely magnetized basalt in the ocean floor adjacent to a spreading centre. Rock magnetized in a direction opposite to the present field is shown shaded. The sea floor cools as it moves away from the ridge axis and the boundary marked *a* is the isotherm at the blocking temperature of the magnetic minerals, at which remanent magnetism is established. Basalt below this level is too hot to have acquired remanence. Boundaries between the normal and reversely magnetized rock mark the blocking temperature isotherms at times when the geomagnetic field reversed.

The tectonic pattern of ridges and trenches is not fixed. They move and old ones are extinguished as new ones form. Perhaps the most obvious new ridge runs down the Red Sea, and a new subduction zone appears to be forming in the Indian Ocean, separating the Australian and Indian plates (Fig. 12.2). Another feature indicative of the transient nature of the tectonic pattern is the formation of back-arc basins. Volcanic

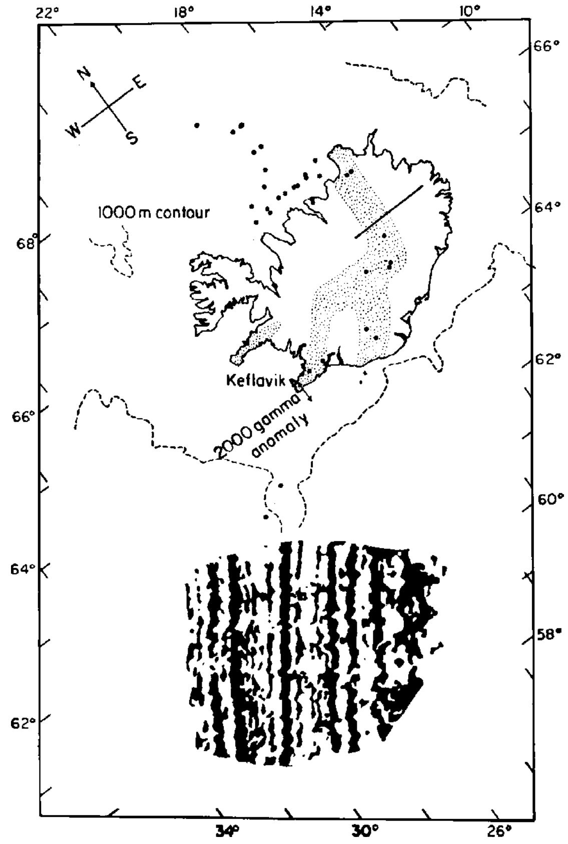


FIGURE 12.10 Linear magnetic anomalies flanking the mid-Atlantic ridge south-west of Iceland, where it is known as the Reykjanes ridge. This shows the surface variation in total field strength, with positive anomalies black. Lighter shading over Iceland shows the area of Quaternary volcanism and dots are earthquake epicentres. Reproduced, by permission, from Heirtzler *et al.* (1966).

island arcs arise where subducted slabs have penetrated to about 100 km depth and are separated from the corresponding trenches by zones referred to as fore-arc basins. As the trenches roll back, tensions develop in the over riding plates on the opposite sides of the arcs causing spreading and the formation of back-arc basins. They are spreading centres that may initiate new ocean ridges and develop into major basins. The Atlantic Ocean is regarded as fully developed back-arc basin. Sea floor topography (Fig. 12.11) shows that in the case of the Mariana back-arc basin a ridge formed right at the arc and split it into two, leaving two fossil arcs. Another apparent example of the same phenomenon is the separation of Baja California from mainland North America.

The ocean rises are elevated because they are hot. As the fresh lithosphere moves away from the ridges it cools and so contracts thermally. Isostatic balance is maintained (Section 20.2), with constant mass in any vertical column, so that the ocean floor deepens systematically away from the ridges (Fig. 20.2). The depth data give valuable evidence of the cooling, because reliable measurements of heat flow are notoriously difficult to make. The principal disturbance is sea water circulation in the crust, especially near the ridge axes. The total contraction between formation of fresh lithosphere at a ridge and subduction is typically 2.1 km. This provides the negative buoyancy that drives subduction and is related to the convective heat transport by the ratio of expansion coefficient to the heat capacity per unit volume, as in Eq. (20.8). Thus, study of the ocean floors has provided the observations that are the basis of the quantitative theory of mantle convection in Chapter 22.

12.4 Plate motions and hot spot traces

The rates of separation of pairs of plates at spreading centres are estimated by the spacing of magnetic anomalies adjacent to the ridges. Directions of motion are not everywhere precisely perpendicular to the ridge axes but are revealed by the orientations of transform faults which separate sections of ridge (Fig. 12.7). Thus, the relative motions of pairs of plates that have common boundaries at spreading centres are reliably determined. Rates of subduction are not as directly observable, although the magnitudes and repetition rates of earthquake displacements appear to give a reasonable indication along at least some zones of rapid convergence (Section 14.7). Further constraint is provided by the condition that spreading and convergence must match along any circular path around the Earth. DeMets *et al.* (1990, 1994) used all the information of this kind to produce a coherent set of plate motions, averaged over a few

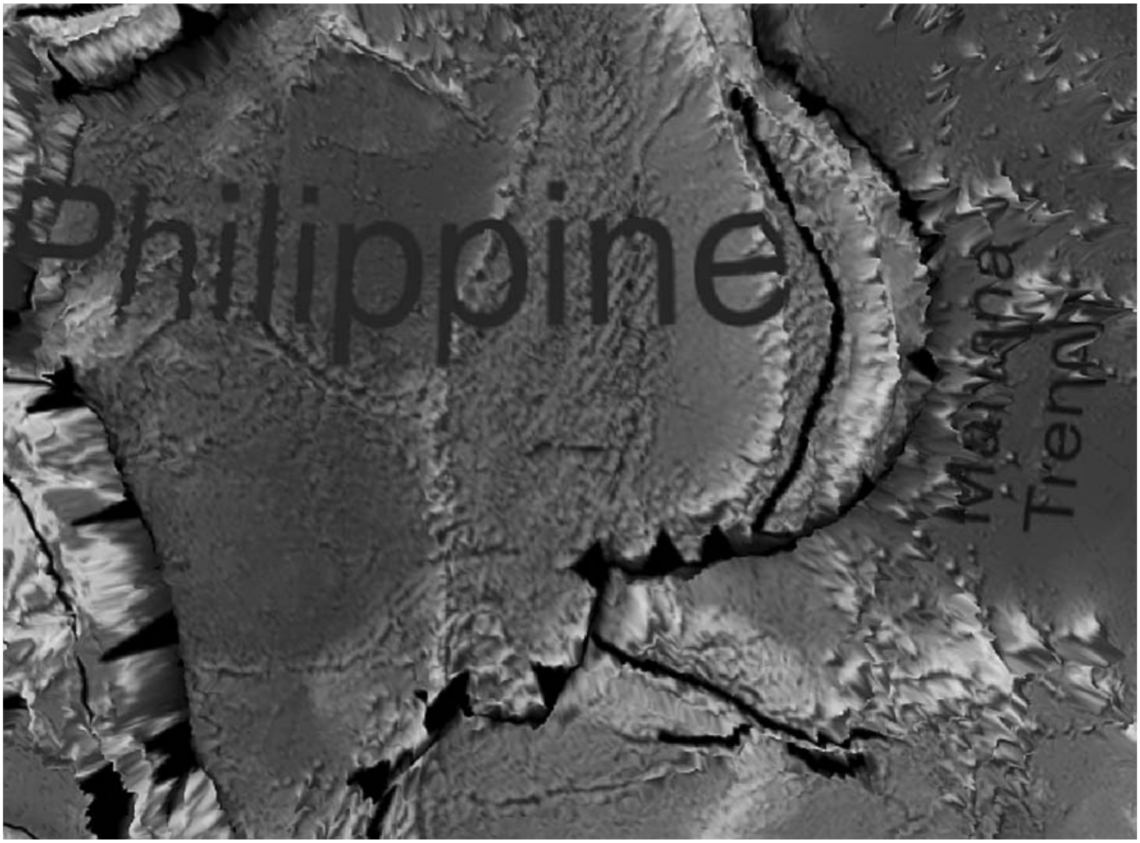


FIGURE 12.11 Sea floor topography in the area of the Mariana island arc. This shows two fossil remnants of an arc which was split by a ridge and a new, active arc parallel to them. Courtesy of David Sandwell.

million years, and Kreemer *et al.* (2003) used an extensive set of data from GPS observations to present a similar set of present day plate motions. Differences between the two solutions are slight enough to be attributable to data uncertainties, demonstrating that plate motion is steady on the time scale of a few million years.

The motion of a rigid plate across the Earth's surface is quantitatively represented as a rotation about an axis through the Earth's centre. The points where the axis cuts the surface are the poles of rotation. The poles are generally remote from the plates, but this is not a necessary condition as rotation of a plate about an axis through itself is conceivable. It is important to note that this gives a complete and general representation of the motion. It is valid for the well-determined relative motions between plates, as well as their less-certain absolute motions. The sign convention established in the pioneering work on rotational mechanics by L. Euler is followed. The pole of rotation is the one about which a plate is moving anticlockwise (either absolutely or relative to another plate) when viewed from above the pole of the motion. The Euler vector is a line from the centre of the Earth to this pole and the magnitude of the vector is the angular speed, ω, of the plate about this axis. Euler rotation vectors add and subtract by the normal vector rules, so that if we write the vector for motion of plate A relative to plate B as ${}_{\mathrm{B}}\boldsymbol{\omega}_{\mathrm{A}}$, then

$$ {}_{\mathrm{C}}\boldsymbol{\omega}_{\mathrm{A}} = {}_{\mathrm{C}}\boldsymbol{\omega}_{\mathrm{B}} + {}_{\mathrm{B}}\boldsymbol{\omega}_{\mathrm{A}}. \tag{12.1} $$

The local speed of the motion, or relative motion, at a point at angular distance θ from the pole of the motion is

$$ v = \omega R \sin\theta, \tag{12.2} $$

where R is the Earth's radius. If the coordinates (latitude, longitude) of the rotational pole (p)

and the point in question (x) are (ϕ_p,λ_p) and (ϕ_x,λ_x) then

$$\cos\theta = \cos\phi_p \cos\phi_x \cos(\lambda_p - \lambda_x) + \sin\phi_p \sin\phi_x. \qquad (12.3)$$

A comprehensive treatment of the geometry of plate motions and how to calculate them is given by Cox and Hart (1986).

Kreemer *et al.* (2003) converted the numerous GPS observations of relative motions between plates to tables of the motions of the individual plates relative to the Pacific plate and relative to a 'no-net-rotation frame', that is, assuming the surface area average to be stationary. They gave some emphasis to the fact that plate boundaries are not sharp but that many of them involve deformation spread over hundreds of kilometers, especially at convergent margins. Nevertheless, most of the surface can be considered divided into plates that are effectively rigid and between which relative motion is unambiguous. An interesting result from this analysis is that the global rms plate velocity is 3.8 cm/year in the no-net-rotation frame. Comprehensive details of plate boundaries and relative motions are presented by Bird (2003).

If the viscous interaction between the plates and the underlying mantle were the same everywhere then the no-net-rotation frame would coincide with the average mantle frame. However, this is not quite true. The asthenosphere appears to be somewhat cooler (or thinner) and more viscous under continents than under oceans and continents tend to 'drag their feet'. The zero net torque condition is not the same as no net rotation, although they may not be very different. There is also a possibility that GPS measurements over a decade or so may not average over sufficient time the relative motions across plate boundaries where displacements occur erratically with major earthquakes at intervals of tens or hundreds of years. Although these are minor quibbles, we consider motion of the plates relative to the underlying mantle to be more fundamental than motion relative to a plate average. In Table 12.1 the geologically derived plate motions from the NUVEL-1 model of DeMets *et al.* (1990) are referred to the position of the Hawaii volcanos as a nominal fixed marker.

Table 12.1 Parameters of the Euler vectors for plate motions relative to the Hawaii hot spot, obtained from motions relative to the Pacific plate by DeMets *et al.* (1990), assuming motion of the Pacific plate given as the first entry. The plates are identified in Fig. 12.2

Plate	Pole latitude ϕ(N)	Pole longitude λ(E)	Angular speed ω (10^{-6} deg/year)
Pacific	−55	144	0.75
Africa	36	−132	0.37
Antarctica	31	−156	0.39
Arabia	68	−26	0.42
Australia	46	52	0.50
Caribbean	12	−138	0.37
Cocos	22	−124	1.72
Eurasia	25	−150	0.40
India	69	−14	0.42
Nazca	41	−122	0.86
N. America	−4	−134	0.35
S. America	−11	−158	0.35
Juan de Fuca	−35	85	0.54
Philippine	−46	−55	0.85

The motion of the India plate relative to the Reunion Island hot spot can be used as a check on the results in Table 12.1. The values in the table for the India plate give its motion relative to the Hawaii hot spot as 4.6 cm/year at 69° east of north. This would also be the motion relative to the Reunion Island hot spot if hot spots were fixed with respect to one another. Reunion Island is on the Africa plate, but the hot spot that produced it began 65 million years ago on the India plate as a massive outpouring of flood basalts in what is now the Deccan area of central India. For about 20 million years the India plate moved northwards across it, leaving the Maldive Islands as the hot spot trace, but there was a change of direction about 40 million years ago when the trace turned north-east. The mid-Indian ridge spreading centre then intervened, so that most of the subsequent trace appears as a somewhat irregular ridge on the Africa plate, extending to the Seychelles. But this does not interfere with our estimate of the motion of

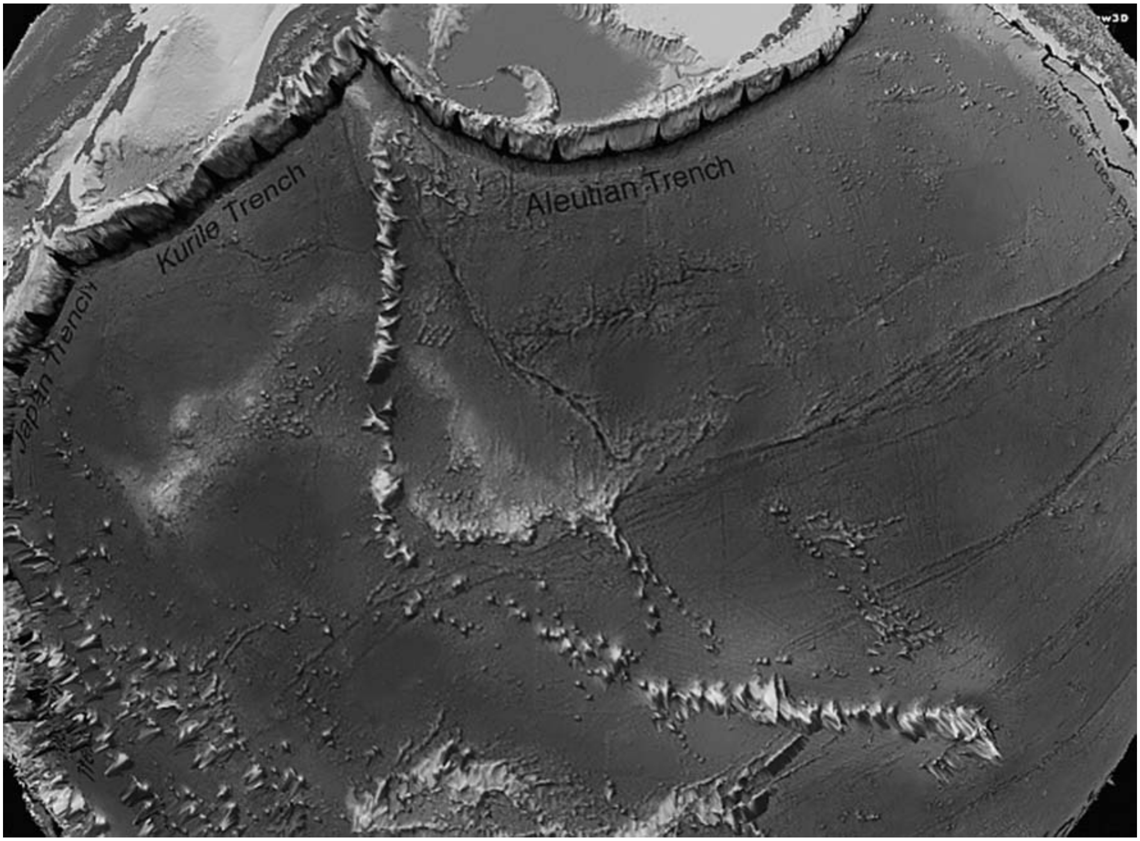


FIGURE 12.12 Trace of the Hawaii hot spot along the chain of the Hawaii Islands and Emperor Seamounts. This is a computer plot of ocean floor topography of the north-central Pacific, with the Japan, Kurile and Aleutian trenches marking the subduction zones at the top of the figure. Courtesy of David Sandwell.

the India plate relative to the hot spot since the direction change, which is seen to be close to 50° east of north at a speed of about 6 cm/year. Thus, the observed motion of the India plate relative to the Reunion Island hot spot differs from its motion relative to the Hawaii hot spot, but the difference is small. Molnar and Atwater (1973) reported a more extensive check on hot spot relative motions, concluding that they had speeds in the range 0.8 to 2.0 cm/year.

The calculation of plate motions relative to the Hawaii hot spot assumes that we know the motion of the Pacific plate. This is obtained from the dated sequence of islands, which give a speed of 8.3 cm/year past the hot spot. For this purpose the island trace, shown in Fig. 12.12, is assumed to follow a great circle, putting the pole of the motion 90° away from the hot spot, at (55° S 14° E), and giving an angular speed of 0.75°/year (the first entry in Table 12.1). An interesting feature of the island chain is the change in direction in the mid-Pacific, apparent in Fig. 12.11. The ages increase linearly from zero at Hawaii Island (19° N 155° W) to 43 million years at (32° N 172° E). At that point there is an abrupt change in the direction of the chain, which continues north as the Emperor Seamounts. The Pacific plate is a major one and a dramatic change in its motion could not have occurred without a reorganization of its boundaries and perhaps other boundaries. Norton (1995) argued that evidence for such reorganization is lacking and attributed the direction change to hot spot motion. This would be incompatible with conventional ideas about plumes, and we have mentioned evidence for a change in the motion of the

India plate about 40 million years ago, so this problem requires further thought.

12.5 The pattern of mantle convection

We summarize by a few key points the observations on which our ideas about mantle convection are based.

(i) Unless the thermal conductivity of the mantle is much higher than we have reason to believe (30 to 100 times that of familiar mantle rocks), diffusive cooling of the deep mantle is insignificant. Without convection, radioactivity would cause the temperature to rise by about 120 K/10^9 years and much more in the distant past.

(ii) The electrical conductivity of the lower mantle is low enough to transmit components of the geomagnetic secular variation with periods as short as a year, disallowing a temperature much higher than that of the upper mantle (Dobson and Brodholt, 2000). This is consistent also with the viscosity variation discussed in the following paragraph. There can be no substantial mid-mantle thermal boundary layer; the mantle must be convecting as a whole and not as separate upper and lower mantle circulations.

(iii) Post-glacial rebound (Section 9.5) indicates a general increase in viscosity with depth below the asthenosphere. Although viscosity is very temperature-sensitive and must be presumed to be locally quite variable, the general trend is consistent with a decrease in convective speed with depth. The extreme viscosity variations occur in the boundary layers, the cool lithosphere, which is quasi-rigid, and the base of the D$''$ layer, which is at core temperature, $\sim$1000 K hotter than the adjacent mantle, with a viscosity lower by at least 10^4. It may even include pockets of partial melt.

(iv) The almost rigid surface plates are moving about with speeds of several centimetres per year. Oceanic crust generated at ridges and disappearing at subduction zones amounts to about 3.4 km^2/year; the oceanic surface area is 3×10^8 km^2, so the average age of the lithosphere at subduction is 90 million years.

(v) Continents move about with their plates, although they appear less mobile than purely oceanic plates, but continental crust is not subducted, being too buoyant (although erosion deposits continental material in the oceans, from which it is subducted). The continental crust has progressively accumulated around ancient cores (shields) that are more than 20 times as old as the oldest ocean floor.

(vi) Isolated volcanic hot spots, of which Hawaii is the most studied example, are not moving with the plates, but more slowly than all but the slowest plates and apparently independently of them. The chemistry and isotopic abundances in the lavas that they produce (ocean island basalt – OIB) are distinct from the mid-ocean ridge basalt (MORB).

(vii) The total surface heat flux from the Earth is 44.2×10^{12} W, of which 8×10^{12} W is attributed to crustal radioactivity, almost all of it in the continents. Disallowing also a core component of the heat loss, which we account for separately, thermal convection of the mantle must be explained by a heat flux of 32×10^{12} W.

(viii) Most of the subduction zones are marked by inclined planes of earthquake foci, in some cases extending to 700 km depth. Seismic tomography indicates that cool subducting slabs penetrate much further, but aseismically. However, they are evidently impeded by the phase transition at 660 km depth.

(ix) Subduction zones are marked also by lines of volcanos, concentrated where the subducting slabs have reached about 100 km depth. Their chemistry includes material derived from sea water.

(x) Vertical motion over most of the mantle at more than a small fraction of the plate speed is disallowed by the sharpness of mantle phase boundaries, especially that at 660 km depth, as pointed out in Section 22.5.

Loper (1985) pointed out that convection is driven by sources of buoyancy (positive or negative) generated at boundaries and we emphasize this in Section 13.1. Both the upper and lower boundaries of the mantle are sources of buoyancy and they drive two quite different modes of convection that are superimposed and at least semi-independent. Plate tectonics is driven by cooling of the lithosphere and the plumes that are responsible for the hot spots carry up heat conducted into the D″ layer from the core. The essential reason for the two different convective styles is that they are controlled by strong but opposite viscosity contrasts. The negatively buoyant lithosphere forms a layer so much more viscous than the underlying mantle that it is virtually rigid and is subducted in broad, coherent slabs, with shearing motion taken up by the asthenosphere and upper mantle. Conversely, the hot material from the base of D″ is readily deformed and self-adjusts to pass up through the mantle with minimum deformation of the surrounding mantle. This means that it forms narrow, axisymmetric plumes, to which the flow is effectively confined. In each case the style of convection is governed by the principle that what happens is what occurs most easily. Deformation is concentrated in the least viscous materials and this situation is reinforced by the heat which is generated by it.

A graphic and convincing illustration of plate tectonic motions was recognized by Duffield (1972) in a natural analogue, the solidifying crust on a convecting lake of lava. Plate spreading centres, subduction zones and transform faults, where plate motions mismatch, were all recorded on a film that shows the motion in 'real time'. The effectiveness of this natural model can be attributed to the fact that the viscosity contrast between the crust and the underlying lava realistically mimics the contrast between the lithosphere and the asthenosphere. Attempts to model mantle convection without this contrast do not include necessary physics. The return flow of asthenospheric material into the ocean ridge spreading centres can be regarded as a passive consequence of subduction-driven convection.

Satisfactory modelling of plumes also requires a high viscosity contrast. Laboratory experiments on liquids heated from below go some way to achieving this if the viscosity is sufficiently temperature dependent and the temperature increment high enough. Viscosity scaling can be brought closer to the terrestrial situation if two different liquids are used, but a model with immiscible fluids does not properly represent the terrestrial situation. But all such models give the same convective pattern, axisymmetric plumes with narrow, rapidly flowing cores. This is inevitable if convection is driven by buoyancy of material with much lower viscosity than the rest of the medium within which it moves.

A feature of plume models is that they are initiated by large, approximately spherical plume heads that melt or soften their way up through the overlying material and are supplied with fresh hot material via narrow conduits. Established mantle plumes, such as those responsible for the hot spots considered in Section 12.4, are simply the narrow conduits of continuing flow, as in Fig. 12.3. When a new plume head reaches the lithosphere it includes partial melt that may appear as massive basalt flows (flood basalts) that completely dwarf any recent volcanism. Courtillot (1999) identified continental areas of such flood basalts, finding seven with dates coinciding with mass extinctions of species (Section 5.5). The best documented is in the Deccan area of India, extending into the ocean and centred near Bombay, from which the hot spot trace leads, via the Maldive Islands, to the centre of current volcanism on Reunion Island, 800 km east of Madagascar. The process by which a new plume head works its way up through the mantle is obviously very slow and on average they have appeared at about 50 million year intervals. So, we suppose that one or several are currently on the way up. Can they be identified with extensive areas of uplift, presumed to be due to heat (superplumes), under Africa and the South Pacific (McNutt, 1998)?

To see why new plumes start and established ones are extinguished, we need to consider the structure and mobility of the D″ layer, where they originate. Core heat is conducted into it, producing a highly mobile layer that is skimmed off into the plumes, but D″ is not uniform. Seismologically, it gives the appearance of having areas where its boundary layer action is fully effective and others where dense material has accumulated, imposing blankets on the core heat. By analogy with the crust we refer to these areas as crypto-oceans and crypto-continents (see Fig. 12.3). The mobile layer moves towards the plume bases and the crypto-continents are carried along with it. When one of them reaches a plume base it cuts off much of the core heat in the vicinity of the plume, starving it and encouraging a new plume to start in a neighbouring crypto-ocean area. The thin, mobile, layer at the base of D″ is less viscous than the mantle well above it by a factor of at least 10^4, but it could be much more than this. Thin, ultra-low velocity zones (ULVZs) at the base of the mantle are recognized seismologically and have been explained in terms of partial melting. However, an alternative explanation is that they are patches of the lowermost mantle where the post-perovskite phase, which is probably the dominant mineral at that level, has absorbed a high proportion of iron from the core (see Sections 17.7 and 19.5).

Plume activity is controlled by the structure and behaviour of D″ and not by the mantle tectonics. The reverse appears not to be true. Courtillot (1999, Chapter 5) traces the evidence that several of the areas of extensive flood basalts mark the initial openings of new ocean basins, implying that continental break-up was triggered by the appearance of massive plume heads. In at least some cases, changes in the plate tectonic pattern, with rearrangement of continental blocks, appear to have been stimulated by the plume activity. But if new plumes arrive at the surface more or less randomly, then they arrive in existing ocean basins at least as often. In this case also they may prompt development of new ridges, but probably rather more easily 'steal' nearby ridges, causing them to shift. The largest single flood basalt province is the submarine Ontong–Java plateau, north-east of the Solomon Islands, with a volume that could be as large as $10^7\,\mathrm{km}^3$. But, unlike continental flood basalts, submarine events are not accompanied by extensive faunal extinctions, presumably because they have little effect on the atmosphere.

The flow of low-viscosity material in the plumes is rapid, at least 1 m/year. The total flow of plume material, estimated to be $40\,\mathrm{km}^3$/year, is much greater than the average global rate of eruption of hot spot basalt, even including occasional new flood basalt provinces. Most of the plume material either underplates the crust or lithosphere, or is injected into the asthenosphere. Sleep (1990) calculated that the elevations of hot spot 'swells', such as that along the Hawaiian chain, are consistent with the buoyancy introduced by the total plume heat flux ($\sim 4 \times 10^{12}\,\mathrm{W}$). It can be regarded as localized contributions to the asthenosphere, but globally it is not a major effect because the rate of accretion of asthenospheric material on to the lithosphere is about 10 times the plume flux.

12.6 Tectonic history and mantle heterogeneity

A current example of the opening of a new ocean basin is the Red Sea, which is the early stage of the separation of Africa from Arabia. There are also areas where convergence of plates appears to have started, but without developed subduction zones; one, marked with a query in Fig. 12.2, occurs in the Indian Ocean, between the Indian and Australian plates. Some new or potential plate boundaries may fail to develop, but others will certainly do so, with the plate tectonic pattern changing on a time scale of tens to hundreds of millions of years. Other boundaries have become inactive. We expect extinct ridges, which are relatively shallow features, to disappear from geological view more rapidly than subduction zones and former plates now deep in the mantle. The durability of remnants of subducted slabs depends on how they are assimilated by the mantle and is a key to a significant tectonic question. It is especially important to consider the lower mantle because convection there is slowest but the heat source is greatest. A mechanism for the distribution of subducted coolness is suggested in the final paragraph of this section.

As mentioned in Section 12.5, without cooling, radioactivity would raise the temperature of the mantle by about $120\,\mathrm{K}/10^9$ years at the present time and much more when radioactivity was stronger. This is superimposed on any temperature changes resulting from adiabatic compression or decompression and occurs whether or not the material is moving. In Section 23.4 it is concluded that long-term cooling of the mantle reduces its potential temperature by about $55\,\mathrm{K}/10^9$ years and in the lower mantle this means a cooling rate of 62 to $88\,\mathrm{K}/10^9$ years, depending on depth, so the total heat loss (at the present rate) is equivalent to cooling by an average rate of $195\,\mathrm{K}/10^9$ years. By this we mean that if a selected volume of the mantle is isolated from convective cooling for 10^9 years while its surroundings are cooled at the average rate, then a temperature contrast of 195 K develops. We consider what this means in terms of tomographically observed heterogeneity (Section 19.7) and convective instability. With the temperature dependences of seismic velocities for the lower mantle in Table 19.1, a 195 K temperature contrast corresponds to a P velocity variation of 0.2% to 0.6% and an S velocity variation of 0.5% to 1.1%. These are at the upper bound of observed variations, so, even though temperature is not a complete explanation of observed heterogeneity, we can use these values to argue that no part of the mantle can be thermally isolated for 10^9 years. It is all convectively cooled.

Convective cooling of the lower mantle presents an interesting problem, bearing in mind that it is more viscous than the upper mantle and its convective motion is slower. The problem is that it must be cooled throughout, although cool lithospheric material reaches it only via thin subduction zones that are separated from one another by thousands of kilometres. Fragmentation of the subducted slabs and wide distribution through the viscous lower mantle material must be discounted as is also thermal diffusion, which is much too slow. The physical scale of thermal diffusion with a relaxation time $\tau = 10^9$ years is indicated by the value of thermal diffusivity, $\eta = 1$ to $2 \times 10^{-6}\,\mathrm{m^2\,s^{-1}}$, giving $(\eta\tau)^{1/2} = 180$ to 250 km.

We can emphasize the nature of the problem by considering the buoyancy of a large volume with a temperature contrast $\Delta T \approx 200\,\mathrm{K}$, as would occur in an isolated volume in 10^9 years, and corresponding density contrast $\Delta\rho/\rho = \alpha\Delta T \approx 3 \times 10^{-3}$, relative to its surroundings. It is convenient to consider a spherical volume, radius $r = 1000\,\mathrm{km}$, because that allows us to use Stokes's law for the viscous drag caused by its motion at speed v through a medium of viscosity η (not to be confused with thermal diffusivity, above),

$$F = 6\pi\eta r v. \tag{12.4}$$

This is equated to the buoyancy force arising from the density contrast,

$$F = (4/3)\pi r^3 \rho\alpha\Delta T g \approx 6.3 \times 10^{20}\,\mathrm{N}, \tag{12.5}$$

with $g = 10\,\mathrm{m\,s^{-2}}$ and $\rho = 5000\,\mathrm{kg\,m^{-3}}$ so that, with $\eta = 10^{22}\,\mathrm{Pa\,s}$ and other numerical values from Appendices F and G,

$$\begin{aligned} v &= (2/9) r^2 \rho\alpha\Delta T g/\eta \approx 3.3 \times 10^{-9}\,\mathrm{m\,s^{-1}} \\ &\approx 10\,\mathrm{cm/year}. \end{aligned} \tag{12.6}$$

This is so much faster than any plausible lower mantle convective speed that the existence of such a large volume with a 200 K density contrast must be discounted. Volumes with this temperature contrast can be no more than a few hundred kilometres in size.

The requirement for subducted coolness to be distributed through the lower mantle on a scale not greater than a few hundred kilometres may appear to present difficulty in devising a plausible convective pattern, but it compels us to recognize that the mantle is not all cooled simultaneously. From time to time the convective pattern changes, so that different parts of the mantle are cooled in turn. New subduction zones (and spreading centres) appear and established ones die out. The intervals between these changes cannot be more than about 100 million years, that is the conventional 'overturn time' of mantle convection. To the extent that tomographically observed irregularities in the lower mantle can be attributed to temperature variations they reflect not just the present convective pattern but relics of earlier patterns that slowly fade. We must presume that the lower mantle is not cooled steadily, but that different regions are cooled in turn, as the convection pattern changes.

13

Convective and tectonic stresses

13.1 Preamble

Decades of general disbelief in continental drift preceded recognition of plate tectonics in the 1960s. In the 1800s, Kelvin noted that the Earth's response to tidal forces indicated an average rigidity modulus exceeding that of steel and, when seismology provided details of internal structure, it showed that the Earth is solid to nearly 2900 km depth. Early proponents of continental drift faced the difficulty that this evidence of solidity appeared incompatible with yielding to any stresses then envisaged. Convection had been contemplated in the 1800s as a means of conveying heat from the deep interior and resolving the age-of-the–Earth problem (Section 4.2). But the idea was stifled, first by difficulty with solar energy and then, when radioactivity was discovered, by recognition that its concentration in continental rocks suggested that the Earth's heat sources were shallow. But by the 1960s, paleomagnetic evidence of continental drift had become overwhelming, mobility of the ocean floors had been recognized and an explanation in terms of convection had become unavoidable. As we now understand, all tectonic processes are ultimately driven by convection. What we see is the surface expression of motion that is necessarily deep.

Thermodynamic arguments (Chapter 22) are central to understanding convective energy and tectonic stresses. The mechanical energy can be derived only by upward transport of heat from very deep sources. This energy is the product of heat flux and a thermodynamic efficiency, giving an estimated power of 7.7×10^{12} W (Section 22.4). In this chapter we consider the implications for tectonic stresses, the point being to demonstrate that the convective power suffices to explain tectonics.

The power generated by mantle convection is dissipated in the mantle and crust. Thermodynamics does not specify how the dissipation is distributed but observations of plate motion give a good measure of the speed of convection and, using this, the calculated power gives an estimate of the stresses involved. The fact that the convective stress, so calculated, is comparable to the stresses apparent in subduction zone earthquakes (Chapter 15) confirms both that convective stresses are responsible for earthquakes and that earthquakes are localized irregularities in the grand pattern of convection (Chapter 12). Thus, the earthquake distribution is a direct indication of that pattern. Moreover, comparison with the gravitational energy released by the negative buoyancy of subducting slabs shows that they provide the principal driving force of convection.

We can make a more general case for this conclusion. As Loper (1985) pointed out in a discussion of the principles of mantle convection, it must be driven by sources of buoyancy (positive or negative) generated at boundaries and our attention must focus on them. We can emphasize this point by imagining a mantle approximating the real one, with an adiabatic temperature gradient and a uniform distribution of heat sources (radioactivity), but supposing it to be thermally isolated from all surroundings so that its temperature rises uniformly. It is large enough for thermal conduction to be ineffective in equalizing the temperature. As the temperature rises, the temperature gradient becomes sub-adiabatic (preventing convection), because the adiabatic gradient is proportional to absolute temperature and to remain adiabatic all temperatures would have to rise proportionately and not uniformly. Internal heat alone does not cause convection, it inhibits convection, and it is only by cooling at the surface that the necessary buoyancy is generated. Of course, heating at the lower boundary (by the core) has the same effect, but if we appeal to heterogeneity of internal sources, then the radioactively well-endowed materials would tend to rise and, if unmixed, stay at the top. Thus, any convection driven by internal sources without reference to surface cooling would have finished early in the life of the Earth.

The conclusion that subduction of cooled lithospheric plates is the essential driving mechanism of convection carries the inference that subduction zones are also areas where dissipation is concentrated. Although this is true, there is no 1:1 correspondence between sources and sinks of convective energy. Subduction does not occur in isolation but is part of a convective cycle with motion throughout the mantle and crust. Stress and energy dissipation occur wherever there is material deformation and consequential stresses extend to areas where there is no noticeable deformation. Observations of crustal stress (Sections 11.5 and 11.6), combined with topography, isostatic balance or unbalance and geological features indicative of ongoing deformation, all contribute to the global picture of a dynamic system.

Our discussion of the mechanism of plate tectonics assumes that it is driven entirely by the loss of heat to the surface from the body of the mantle. The convective plumes driven by core heat are geometrically very different from plate-driven convection and operate apparently independently of it. However, plumes influence plate motion. They contribute to ‘ridge push’, as in Iceland, and probably aid the initiation of new spreading centres, as in East Africa. To reach the surface, core heat must traverse the entire depth of the mantle, giving the plumes a high thermodynamic efficiency (39% by Fig. 22.5).

Stresses within the lithosphere are required to support topography. In Chapter 9 we examine large-scale loading of the Earth’s surface by ice sheets that cause relaxation in the mantle towards isostatic equilibrium. The relaxation is rapid compared with geological processes because the material deformation is quite small. Similarly, topography with scale lengths greater than a few hundred kilometres is in isostatic equilibrium. However, at shorter wavelengths, loads are supported by elastic flexure of the lithosphere, which indicates that the lithosphere can retain elastic strains over geologic time and so must have an effective viscosity several orders of magnitude higher than that of the underlying mantle. A high-floating lithospheric block, such as a plateau, has internal stresses that would spread it out to sea level were it not held together elastically. On the other hand low-lying blocks are in compression. The stress state can be calculated by integrating over boundary forces, and superimposing the body force of gravity. For an internally homogeneous fluid planet with blocks floating on its surface the difference between vertical and horizontal stresses in the blocks is linearly related to variations in geoid height. Regions of positive geoid correspond to a state of extension and negative to compression. While the Earth is more complicated because of internal heterogeneity, an overall correspondence between geoid height and stress state can be recognized.

The very long wavelength features of the geoid are attributed to density variations in the lower mantle and are believed to indicate residua of past subduction. These effects must be subtracted before the geoid features related to lithospheric structure can be recognized. Only the long wavelength density variations in the lower mantle are gravitationally apparent at the surface as the higher harmonic terms are geometrically attenuated. The geoid features of intermediate wavelengths can be interpreted in terms of topography and density variations in an isostatically balanced lithosphere–asthenosphere system. A plot of the geoid, with harmonic degrees of six and less subtracted (Fig. 13.1(b)), relates to details in the stress map (Fig. 11.11).

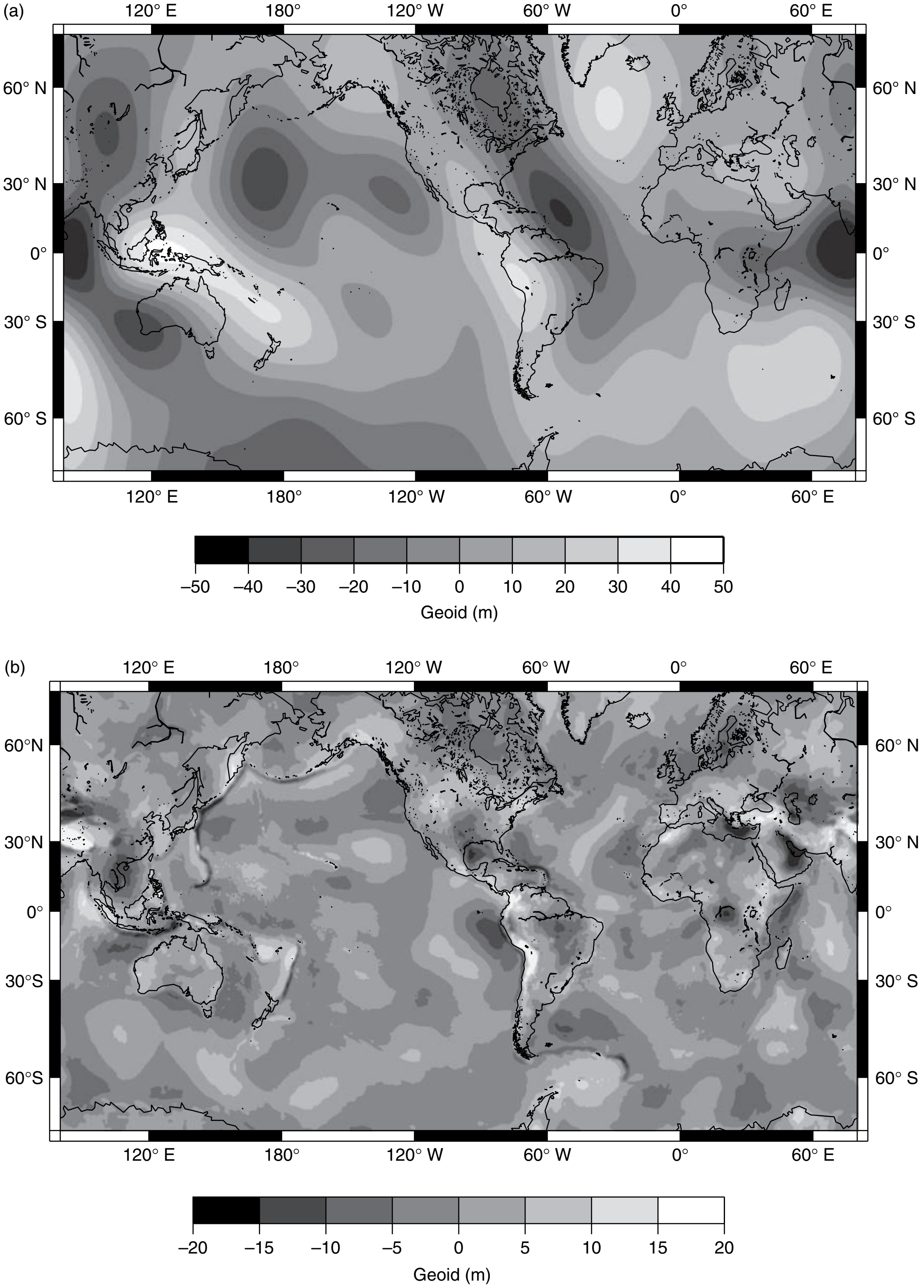


FIGURE 13.1 Filtered versions of Geoid EGM96 by Lemoine *et al.* (1998) selecting (a) wavelengths from 5000 km to 15 000 km and (b) wavelengths from 100 km to 5000 km. Figures by King (2002). In (b) East Africa is an area of positive geoid anomaly, consistent with extensional stress. Other elevated regions in extension include the Tibetan plateau, the Basin and Range of west North America, and the Andes.

Regions of high elevation, plateaux and mountain ranges, are in tension and correspond to positive anomalies in the filtered geoid. The most dramatic effect of filtering is in seen in the plateau of East Africa, where the world's largest continental rift is found. Without harmonic filtering East Africa is in a geoid low, but when low harmonics are subtracted it appears in a high.

Topography is built up through inelastic processes, such as folding and earthquakes. A particular case occurs above subduction zones in the material that accumulates above subducting plates, as the so-called accretionary wedges. The subduction acts as a conveyor belt, piling sedimentary material onto the over-riding plates that act as backstops. The material fails in a brittle fashion and adjusts to a wedge shape that is controlled by friction between broken surfaces with internal fluid pressure. The wedge angle is referred to as a critical taper. This is modelled by a theory that balances the gravitational and frictional forces that are calculated in terms of the topographic slope and dip of the sliding basement. It applies equally well to large-scale features, such as the Himalaya and Taiwan, and to small-scale models such as sand box experiments.

13.2 Convective energy, stress and mantle viscosity

The mechanical energy of mantle convection is estimated in Chapter 22 from a fundamental thermodynamic argument. For the whole mantle the power is 7.7×10^{12} W. It may seem surprising that, of this total, only 2.4×10^{12} W is produced in the lower mantle and 5.3×10^{12} W in the upper mantle. The reason is that the lower mantle heat is distributed and much of it is transported over a limited temperature range, whereas all of the lower mantle heat traverses the entire upper mantle. The significance of these numbers is that they are values of power that must be used in deforming mantle (and crustal) material and not merely power that is available in principle. They must be equated to integrals of the product of stress and strain rate through the whole volume of the mantle. For the upper mantle we have direct evidence of stress and strain rate and can demonstrate that the thermodynamically calculated energy suffices to explain them. It is not so clear what is happening in the lower mantle but, as these numbers show, the power generation per unit volume is more than five times smaller than in the upper mantle. While we cannot require that dissipation be distributed in the same way, it is inevitable that convection is more vigorous in the upper mantle.

In making a quantitative assessment of the relationship between convective energy and stress, we use details of plate geometries and speeds from a comprehensive survey by Bird (2003) and especially his Table 3. The rate of production of new crust is 3.36 km^2/year, more than 90% of it at the ocean ridges, with a small contribution by their extensions on to the continents. Concentrating attention on the 67 000 km of oceanic spreading ridges, the average spreading rate is 5.1 cm/year or 2.5 cm/year ($7.4 \times 10^{-10}\,\mathrm{m\,s^{-1}}$) each way from the ridge axes. By assuming this to be the average speed, v, of the surface plates across the underlying mantle, with total area equal to the surface area of the Earth ($A = 5.1 \times 10^{14}\,\mathrm{m^2}$), we can write the viscous dissipation by this motion, $\dot{E}_\mathrm{P}$, as a fraction f_p of the upper mantle convective energy, $\dot{E} = 5.3 \times 10^{12}$ W,

$$\dot{E}_\mathrm{P} = f_\mathrm{P}\dot{E} = Av\sigma, \tag{13.1}$$

where σ is the shear stress between the plate and the mantle below. The deformation is concentrated in the layer of lowest viscosity (the asthenosphere). Let this have an effective thickness H and viscosity η. Then the strain rate is

$$\dot{\varepsilon} = v/H \tag{13.2}$$

and the viscosity is

$$\eta = \sigma/\dot{\varepsilon} = \sigma H/v. \tag{13.3}$$

We have independent evidence of the relationship between η and H from post- glacial rebound (Section 9.5). Writing $H = n \times 100$ km, Eq. (9.35) gives

$$\eta\,(\mathrm{Pa\,s}) = 6.5 \times 10^{18}\,n^3. \tag{13.4}$$

Identifying this with Eq. (13.3) and substituting the value of v ($7.4 \times 10^{-10}\,\mathrm{m\,s^{-1}}$), we have

$$\sigma\,(\mathrm{Pa}) = 4.8 \times 10^4 n^2. \tag{13.5}$$

If we assume $n = 2$, that is a 200 km thick asthenosphere, which we regard as the upper limit of the plausible range, then $\sigma = 1.9 \times 10^5$ Pa. This is much less than the stress release typical of large earthquakes, 10^7 Pa or so, and indicates that the plates slide relatively freely over the asthenosphere. Taking $n = 1.5$, that is a 150 km average asthenospheric thickness, $\sigma = 1.1 \times 10^5$ Pa and, by Eq. (13.1), the viscous dissipation is 4.1×10^{10} W with $f_p = 0.0077$. The corresponding average asthenospheric viscosity is 2.2×10^{19} Pa s. Although there are obvious uncertainties in this calculation it suffices to indicate that the surface motion of the plates is little more than a passive consequence of mantle convection. An explanation of the lithospheric stress in terms of the gravitational potential energy of ridge elevation is given in Section 13.4.

Now consider the energy of subduction. The process becomes somewhat confused at 660 km depth so we concentrate initially on the upper mantle, that is we calculate the gravitational energy of subduction to this depth of $3.36\,\mathrm{km^2}$/year of lithosphere (assuming that it all reaches that depth). Its negative buoyancy is due to the thermal shrinkage apparent from ocean depth observations, as discussed in Section 20.2. With the allowance for the isostatic component of ocean floor deepening we take the average shrinkage to be 2.1 km, of which $\Delta z_c = 200$ m occurs in the igneous crust, density $\rho_c = 2900\,\mathrm{kg\,m^{-3}}$, leaving $\Delta z_m = 1900$ m to be accounted for by shrinkage of mantle material, with density $\rho_m = 3370\,\mathrm{kg\,m^{-3}}$. The thermal expansion coefficient, α, decreases with depth, being only 0.7 of the zero pressure value at 660 km, and the density increases by the factor 1.18 over the same range (including effects of the phase transitions above 660 km). Thus the product $(\alpha\rho)$ decreases by the factor 0.83 over this depth range and we take the average of this product in the upper mantle to be smaller than the surface value by $\langle\alpha\rho\rangle/(\alpha\rho)_0 = 0.915$. The negative buoyancy arising from shrinkage is multiplied by this factor. It is partly offset by the positive buoyancy of the intrinsic density difference, $(\rho_m - \rho_c) = 470\,\mathrm{kg\,m^{-3}}$, between the crustal component, thickness $z_c = 7$ km, and the mantle into which it sinks. With these numbers the net downward buoyancy force per square metre of oceanic lithosphere is

$$F/A = g[(\rho_m \Delta z_m + \rho_c \Delta z_c)\langle\alpha\rho\rangle/(\alpha\rho)_0 - (\rho_m - \rho_c)z_c] = 5.66 \times 10^7\,\mathrm{N\,m^{-2}}, \tag{13.6}$$

where $g = 9.92\,\mathrm{m\,s^{-2}}$ is the average gravity over the 660 km depth of subduction. With $3.36\,\mathrm{km^2/year} = 0.106\,\mathrm{m^2\,s^{-1}}$ of lithosphere descending 6.6×10^5 m, the rate of gravitational energy release is 4.0×10^{12} W, more than 70% of the upper mantle energy estimated thermodynamically in Section 22.4. If we assume that a significant fraction of the crustal component escapes full subduction, then the energy release by subduction is even nearer to the thermodynamic result, but there is still a difference to be accounted for by convective forces elsewhere in the upper mantle. The allowance for crustal buoyancy introduces the problem that basaltic oceanic crust is believed to convert to a denser form, termed eclogite, at depth, removing its buoyancy, with an effect similar to the assumption that the crust does not subduct. However, this is a phase transition and needs special consideration in the calculation of convective energy (Section 22.3). If we remove the crustal term in Eq. (13.6), then we must allow for a compensating energy term elsewhere in the convective cycle.

We estimate the mantle viscosity by equating the gravitational energy release by subduction to the viscous dissipation by subducting slabs. Their effective surface area is the 51 000 km length of the subduction zones multiplied by the 660 km/sin θ down-slab distance, where θ is the dip angle, which we take as 45°, all multiplied by 2 because the slabs experience drag on both top and bottom surfaces, giving a total area $A_S = 9.5 \times 10^{13}\,\mathrm{m^2}$. Since the total surface area is conserved and the $3.36\,\mathrm{km^2}$/year of crustal generation is matched by the same rate of subduction, the average subduction speed is 6.6 cm/year, that is $v = 2.1 \times 10^{-9}\,\mathrm{m\,s^{-1}}$. We identify the energy dissipation with the gravitational energy release estimated above, so that

$$\dot{E} = A_s v \sigma = 4.0 \times 10^{12}\,\mathrm{W}, \tag{13.7}$$

where σ is the consequent mantle stress, for which this argument gives 2.0×10^7 Pa. This is at the high end of the range of subduction zone stresses inferred from earthquakes and confirms the adequacy of convection to maintain tectonic activity, including earthquakes.

In applying Eq. (13.3) to estimate the viscosity of the mantle as a whole from its resistance to slab subduction, there is no obvious value of H to assume. Although the shearing action may be spread quite widely, the self-softening effect of the dissipation will tend to concentrate it. If we assume $H = 150$ km, with $v = 2.1 \times 10^{-9}\,\mathrm{m\,s^{-1}}$, as above, then $\eta = 1.4 \times 10^{21}$ Pa s. Acknowledging the wide variations in viscosity, this must be regarded as agreement with the post-glacial rebound estimate of upper mantle viscosity (Section 9.5).

So far we have considered only surface motion of the plates and subduction in the upper mantle, essentially because these are the observed features of mantle convection. Discussion of the pattern of return flow and the involvement of the lower mantle is necessarily more speculative, but there are several considerations that provide constraints on hypotheses.

(i) The adiabatic temperature gradient is proportional to the absolute temperature, T, as in Eqs. (19.55) and (19.56), so, assuming the mantle to remain approximately adiabatic as it cools, the temperature drop, and therefore the heat lost by any material element is proportional to T. The high temperature of the lower mantle means that the heat lost by the lower mantle is a greater proportion of the total heat loss (77%) than its mass is of the total mass (73%). The 73% factor applies to radiogenic heat, which we assume to be distributed according to density. Core heat is treated separately.

(ii) The mechanical power generated in the lower mantle is about 30% of that generated in the mantle as a whole, although it comprises 2/3 of the volume. Convective energy is calculated thermodynamically, but in relating it to buoyancy forces, it is necessary to recognize that the thermal expansion coefficient decreases by a factor of three over the depth range of the mantle. Also the temperature ratios are lower in the lower mantle than in the upper mantle, reducing the Carnot efficiency of convection.

(iii) Viscosity of the lower mantle is higher than that of the upper mantle, perhaps by a factor of about ten.

(iv) At the present rate of subduction a volume equal to the upper mantle is cycled through the lithosphere in one billion years or less. For a volume equal to the whole mantle the corresponding time is three billion years.

(v) Two thirds of the mantle heat lost to the surface is accounted for by radioactivity, which is distributed throughout the mantle. Unless we allow the implausible hypothesis that parts of the mantle are consistently heating up over billions of years, subducted coolness must, by some mechanism, be distributed throughout the volume.

(vi) Except in subduction zones, mantle phase boundaries are observed to be seismically sharp and essentially uniform in depth. As shown in Section 22.4, this restricts the speed of upward return flow in the upper mantle to a small fraction of the speed of subduction ($<10\%$), requiring it to be very broad in scale and probably mantle-wide.

(vii) The temperature of the lower mantle inferred from its electrical conductivity is not dramatically higher than that of the upper mantle (Dobson and Brodholt, 2000), disallowing the hypothesis of a mid-mantle thermal boundary between separate upper and lower mantle circulations.

(viii) At least some of the slab material maintains its coherence in penetrating the lower mantle, but the phase transition at 660 km is an impediment to convection.

(ix) The thermodynamically calculated convective energy disallows the idea that all of the subducted material eventually reaches the bottom of the mantle.

(x) Mantle-derived rocks indicate that isotopically distinct source regions have survived convective mixing.

A convection model satisfying all of these conditions is demanding of ingenuity. With the very limited direct evidence of just how the lower mantle is convecting, it is not surprising that there are divergent ideas. We can make the general observation that, since the heat flux per unit area at radius r decreases with depth (decreasing r), then so does the speed of convection. This is qualitatively consistent with increasing viscosity, but an increase by a factor of order ten, as suggested by post-glacial rebound, can be accommodated only by assuming that convective motion is much less concentrated than in the upper mantle. We need to note also that the assumption that the mantle heat loss is derived from throughout the mantle is equivalent to the assumption that the subducted coolness is also distributed throughout the lower mantle (although not necessarily steadily). Rather little reaches the base of the mantle but two thirds passes through 660 km depth. This makes it easier to understand that lower mantle convection is slower, but it is not as easy to see just how the coolness is distributed through a medium with the lower mantle viscosity.

We can reasonably ask: is it necessary to assume that convection is the only significant mechanism for mantle cooling? The alternative is conduction, but the conductivity would need to be about $80\,\mathrm{W\,m^{-1}\,K^{-1}}$ and the only possibility for this would be radiative conduction, as has sometimes been suggested. Applying Eq. (19.62) we see that this would mean an opacity $\varepsilon \approx 150\,\mathrm{m^{-1}}$, corresponding to an optical depth of 7 mm. It appears implausible that iron-bearing lower mantle minerals could be so transparent. In any case, if the lower mantle conductivity were as high as this then the thermal structure would be quite different from our present understanding. Core heat would be conducted up through the lower mantle with no D″ thermal boundary layer, or supply of hot material for deep mantle plumes, eliminating the explanation of the mismatch of core and mantle temperatures. Convection would be weak, if it occurred at all, with negligible power for tectonics. So, we require bodily convection of the whole mantle, and lack of a thermal boundary layer at 660 km depth means that the lower mantle cannot be considered separately. Observation (vi) above means that, since the upward return flow in the upper mantle is very slow and broad in scale, it is inevitable that this motion extends into the lower mantle. These conclusions alert us to the problem that the lower mantle is cooled by subducted material rapidly enough to ensure that there are no persistent large-scale temperature variations that would cause implausibly large buoyancy forces. In the final paragraph of Section 12.6 we suggest that the logical explanation is that the convective pattern repeatedly changes, on a time scale much shorter than 10^9 years, so that different regions of the lower mantle are cooled sequentially and not simultaneously.

13.3 Buoyancy forces in deep mantle plumes

The contrast in temperature between the lower mantle, extrapolated to the core–mantle boundary from the 660 km deep phase transition, and the outer core, extrapolated from the solidification point at the inner core boundary, is at least several hundred degrees and probably nearer to 1000 K. This is the temperature increment in the thermal boundary layer at the base of the mantle. It is identified as the D″ layer, although there are also compositional and phase heterogeneities that complicate the interpretation of D″. Since viscosity is a strong function of temperature, the mantle material adjacent to the core, being at core temperature, is much less viscous than the material higher up. We can use Eq. (10.27) to show how big the effect is. Taking $n = 1$ for linear (Newtonian) viscosity, $\eta = \sigma/\dot{\varepsilon}$, the ratio of viscosities at temperatures T_1 and T_2 is

$$\eta_1/\eta_2 = \exp[gT_{\mathrm{M}}(1/T_1 - 1/T_2)]. \qquad (13.8)$$

We cannot choose plausible values of T_1, T_2, T_{M} that give a value of this ratio smaller than 10^4:1.

The ratio is smallest if the higher temperature, T_2, is near to the solidus temperature, T_M. Taking $T_1 = 2785\,\mathrm{K}$, $T_2 = 3739\,\mathrm{K}$ from Appendix G with $T_M = 4000\,\mathrm{K}$, $\eta_1/\eta_2 = 6 \times 10^4$. Variations of 200 K in the assumed value of T_M change this ratio by a factor less than two. This is the factor by which the material immediately adjacent to the core is softened relative to the bulk of the lower mantle.

A thin layer of the most softened material at the bottom of the mantle (the base of D″), being thermally dilated and strongly buoyant, feeds the deep mantle plumes. Its reduced viscosity explains why the plumes can be narrow and fast flowing. Not all of the plume material can be at the boundary temperature and its average viscosity is a compromise with the viscosity of material drawn from slightly higher up in D″. But the softest material flows fastest and the flow in D″ is restricted to a very thin layer of the lowest viscosity at the bottom and involves only a very small fraction of the boundary layer. It is convenient to refer to an average viscosity, which cannot plausibly exceed 10^{-4} of the general lower mantle viscosity. Taking this to be 10^{22} Pa s, we have an average plume viscosity of 10^{18} Pa s. The dynamics of plumes transporting this softened material upwards through the mantle are presented by Loper and Stacey (1983) for Newtonian viscosity ($n = 1$ in Eq. 10.27) and generalized to non-Newtonian viscosity by Loper (1984). The material is softest at the centre, where the flow is concentrated, and the buoyancy gives it a reduced pressure, causing a slow inward collapse of the mantle, which restricts the diameter and throttles the flow in a self-stabilizing way. Plumes have thermal halos with diameters of order 10 times their effective fluid diameters. Here we consider a simplified version, a vertical pipe of fixed radius, a, conveying 'fluid' of uniform viscosity, 10^{18} Pa s.

If there are 20 mantle plumes conveying altogether $4 \times 10^{12}\,\mathrm{W}$ of core heat to the surface, or to the asthenosphere, that is $2 \times 10^{11}\,\mathrm{W}$ each, and, by Fig. 22.5, 39% of this is converted to mechanical energy, then the mechanical power of the buoyancy of each plume is $E = 0.8 \times 10^{11}\,\mathrm{W}$. This is $E/l = 2.4 \times 10^4\,\mathrm{W}$ per metre of plume, assuming a vertical rise, so that if the volume rate of flow is $\dot{V} = \pi a^2 \bar{v}$ for mean speed $\bar{v}$ the buoyancy force per metre of plume is

$$F/l = (\dot{E}/l)/\bar{v} = (\dot{E}/l)\pi a^2/\dot{V}. \tag{13.9}$$

But $\dot{V}$ is given by the heat transported,

$$\dot{Q} = \dot{V}\rho C_P \Delta T = 2 \times 10^{11}\mathrm{W}. \tag{13.10}$$

With $\rho = 4500\,\mathrm{kg\,m^{-3}}$ averaged over the mantle depth, $C_P = 1200\,\mathrm{J\,K^{-1}\,kg^{-1}}$ and taking $\Delta T = 800\,\mathrm{K}$ (slightly less than $(T_2 - T_1)$ by the plume averaging calculation of Stacey and Loper (1983)), we have $\dot{V} = 46\,\mathrm{m^3\,s^{-1}}$. Then by Eq. (13.9),

$$F/l(\mathrm{Nm^{-1}}) = 1900\,a^2\ (a\,\text{in metres}). \tag{13.11}$$

This force is balanced by the viscous drag, for which we adopt Poiseuille's formula for the volume rate of flow driven by a pressure difference ΔP over length l,

$$\dot{V} = \pi \Delta P a^4/8\eta l = (F/l)a^2/8\eta, \tag{13.12}$$

in which we have substituted $\pi a^2 \Delta P = F$. With (F/l) given by Eq. (13.11) and taking $\eta = 10^{18}$ Pa s, we have

$$a = (8\,\dot{V}\,\eta/1900)^{1/4} = 21\,\mathrm{km}. \tag{13.13}$$

There are obvious uncertainties and approximations in this calculation but the ¼ power in Eq. (13.13) ensures that the calculated plume radius is not very sensitive to them. The plumes are narrow, although the estimated radius refers only to the very hot flowing channels which are surrounded by much wider thermal halos that may be seen by seismological observations. The average speed of plume material,

$$\bar{v} = \dot{V}/\pi a^2 = 3.3 \times 10^{-8}\,\mathrm{m\,s^{-1}} = 1.0\,\mathrm{m/year}. \tag{13.14}$$

is much greater than the plate speed. These calculations ignore the effect of partial melting which further reduces viscosity and plume radius and increases the speed.

13.4 Topographic stress

The lithosphere is thought to be strong enough to sustain some level of deviatoric stress for at least 10^8 years. Deviatoric stresses relax in

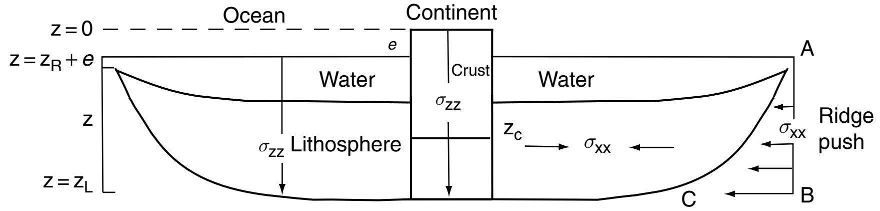


FIGURE 13.2 A schematic representation of the Africa plate with ridges on either side. The symmetry simplifies the calculation of stresses because we assume it to be stationary, with no asthenospheric drag. At a mid-ocean ridge the asthenosphere exerts pressure on the lithosphere and this is resolved to give the horizontal stress. The integral of the resolved stress over the boundary is called ridge push. Vertical stresses are attributed to gravity.

the weaker asthenosphere, where the dominant stress is pressure. The pressure of elevated asthenosphere at mid-ocean ridges exerts a horizontal force on the oceanic lithosphere. 'Ridge push' is the horizontal component of the distributed force from the asthenospheric pressure across the thickening lithosphere. It compresses the plates. Plate acceleration is insignificant, so ridge push is balanced by resistive forces, oppositely directed ridge push from distant ridges, viscous drag by the asthenosphere, friction on transform faults and forces from adjacent plates or subduction zones. Of the forces that act on the lithosphere, calculation of ridge push is most straightforward. Figure 13.2 shows the situation of the Africa plate, which is geometrically symmetrical, with ridges both east and west, and is believed to be nearly stationary with respect to the underlying mantle. This allows us to neglect any viscous drag of the asthenosphere, or the effect of subduction zones, simplifying the calculation of internal stresses. The ridges are assumed to be spreading symmetrically, so that they are retreating from the continent and leaving fresh oceanic lithosphere essentially stationary with respect to the underlying mantle.

The initially uplifted lithosphere thickens as it subsides, and ridge push is the force per unit length of ridge, given by the line integral of the horizontal component of the asthenospheric pressure on the underside of the lithosphere between points A and C in Fig. 13.2,

$$RP = \int_{AC} P_a \mathrm{d}l \cos(\theta) = \int_{AB} P_a \mathrm{d}z, \qquad (13.15)$$

where P_a is the asthenospheric pressure, dl is the differential line length along AC and θ is the angle dl makes with the z axis, taken as downwards (note that in this chapter we adopt the rock mechanics sign convention that compressive stress is positive). This force is balanced by the stress within the lithosphere, so that the total integrated force on the lithosphere can be calculated as the line integral of P_a over AB, avoiding the need to integrate over the curved surface. In the Africa case, ridge push simply imposes a static compressive stress on the lithosphere, but in other situations, with the lithosphere moving across the asthenosphere, the ridge push is distributed across the lithosphere–asthenosphere boundary and overcomes the viscous drag. Noting that the dimension AB in Fig. 13.2 is about 0.02 of the horizontal extent of a lithospheric plate (~5000 km), the average shear stress between the lithosphere and asthenosphere is ~0.02 of the ridge push calculated for the Africa case. If this stress is assumed to be sufficient to overcome the viscous drag, then the asthenospheric viscosity does not exceed 4×10^{19} Pa s. We use the ridge as a reference in calculating stresses elsewhere in the lithosphere.

To examine the stresses that develop in uplifted regions in isostatic equilibrium, consider a block of floating material with its upper surface a distance e above the fluid surface, as in Fig. 13.3. For the analysis that follows, its lateral extent is assumed to be much greater than its thickness. Isostasy requires that at the depth of compensation, at the bottom of the block, the vertical stress in the block and pressure in the fluid are the same. However, elsewhere

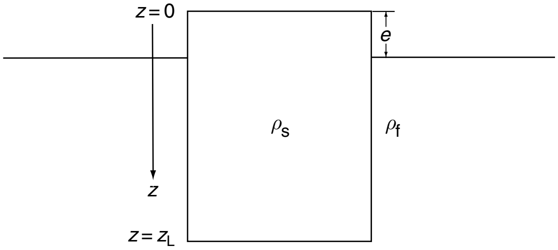


FIGURE 13.3 Geometry for calculation of stresses in an isostatically balanced block.

the vertical and horizontal stresses within the block are different. We shall show that the vertical stress is larger, on average, by $(1/2)\rho_s ge$, where ρ_s is the density of the solid, and this difference is proportional to the difference in the geoid height in the block and fluid areas. Let z be measured downwards from the solid surface (Fig 13.3). The horizontal pressure exerted by the fluid is

$$\begin{aligned} \sigma_{xx} &= \rho_f g(z - e), \ z \geq e \\ &= 0 \qquad\qquad z < e \end{aligned} \tag{13.16}$$

(recalling the convention that compression is taken as positive). The vertical stress in the solid is

$$\sigma_{zz} = \rho_s gz. \tag{13.17}$$

The average vertical stress is found by integrating Eq. (13.17) from 0 to the compensation depth, z_L, and dividing by z_L,

$$\bar{\sigma}_{zz} = \frac{1}{2}\rho_s gz_L. \tag{13.18}$$

The average horizontal stress is found by integrating Eq. (13.16) from e to z_L and dividing by z_L,

$$\bar{\sigma}_{xx} = \frac{1}{2}\rho_f g\frac{(z_L - e)^2}{z_L}. \tag{13.19}$$

Combining Eq. (13.19) with the isostatic relation (for uniform gravity), $z_L\rho_s = (z_L - e)\rho_f$, the average stress difference is

$$\bar{\sigma}_{zz} - \bar{\sigma}_{xx} = \frac{1}{2}\rho_s ge. \tag{13.20}$$

This stress difference acts in a manner that would cause the block to spread out. Such a deviatoric stress, with the maximum principal stress vertical, is the Anderson condition for normal faulting (Section 11.5).

We have developed this result for uniform densities. However, it is true for arbitrary depth variation of density in regions of isostatic equilibrium. Consider two regions in isostatic equilibrium with density variation $\rho_1(z)$ and $\rho_2(z)$, joined at a common vertical boundary, with depth of compensation D. Let region 2 be fluid-like, with no deviatoric stress. We can calculate the average vertical stresses in region 1 due to the vertical gravity body force and the average and horizontal stresses from the horizontal traction applied by region 2,

$$\bar{\sigma}_{zz} = \frac{1}{D}\int_0^D\int_0^z \rho_1(z')gdz'dz,$$

$$\bar{\sigma}_{xx} = \frac{1}{D}\int_0^D\int_0^z \rho_2(z')gdz'dz. \tag{13.21}$$

Integrating (13.21) by parts,

$$\bar{\sigma}_{zz} = \int_0^D \rho_1(z)gdz - \frac{1}{D}\int_0^D z\rho_1(z)gdz$$

$$\bar{\sigma}_{xx} = \int_0^D \rho_2(z)gdz - \frac{1}{D}\int_0^D z\rho_2(z)gdz. \tag{13.22}$$

The average stress difference is

$$\begin{aligned}(\bar{\sigma}_{zz} - \bar{\sigma}_{xx}) = &\int_0^D \rho_1(z)gdz - \frac{1}{D}\int_0^D z\rho_1(z)gdz \\ &- \int_0^D \rho_2(z)gdz + \frac{1}{D}\int_0^D z\rho_2(z)gdz.\end{aligned} \tag{13.23}$$

Isostatic balance requires

$$\int_0^D \rho_1(z)gdz = \int_0^D \rho_2(z)gdz, \tag{13.24}$$

so that Eq. (13.23) becomes

$$(\bar{\sigma}_{zz} - \bar{\sigma}_{xx}) = \Delta\sigma = \frac{1}{D}\int_0^D z\Delta\rho(z)gdz, \tag{13.25}$$

where $\Delta\rho(z) = \rho_2(z) - \rho_1(z)$.

As in Eq. (9.20), the difference in geoid height between the regions is given by

$$\Delta N = -\frac{2\pi G}{g}\int_0^D [\rho_1(z) - \rho_1(z)]z\mathrm{d}z \tag{13.26}$$

where G is the gravitational constant. Comparing Eqs. (13.25) and (13.26) we see that, in general,

$$\Delta\sigma = \frac{g^2}{2\pi GD}\Delta N, \tag{13.27}$$

with g assumed to be the same in both regions. In areas where the geoid is high compared with surrounding areas, the vertical stress is greater, leading to normal faulting. The discussion here considers a first-order treatment that suffices for an understanding of tectonic stresses. For some purposes higher orders may need to be considered (Chambat and Valette, 2005).

Figure 13.1(a) is a plot of the geoid filtered to select long wavelengths (5000 km $< \lambda <$ 15 000 km) and Fig. 13.1(b) shows the higher spatial frequencies (100 km $< \lambda <$ 5000 km), with the long wavelengths filtered off. The long wavelength geoid (harmonic degrees ≤ 6) is attributed to features of the deep mantle (Hager and Richards, 1989), especially the lower mantle where the viscosity is high (Section 9.5), and changes caused by convection are necessarily slow. Richards and Engbretson (1992) found a strong correlation of the long-wavelength features with the pattern of deep mantle densities expected by extrapolating back for 100 million years the present pattern of slab subduction. This is evidence that slabs penetrated, in at least some cases, to the base of the mantle and this is confirmed by the correlation with tomographic models. The fact that the correlation is limited to the hundred million year extrapolation is consistent with the convection model discussed in Section 12.6, which requires the pattern of convective motion to change at intervals of this order, so that surviving pieces of earlier remnant slabs would not be correlated with the present pattern, although they would take much longer than this to decay by thermal diffusion. (The thermal halo of an old slab would not be distinguishable in observations of long wavelengths from a young, thin slab with the same integrated coolness).

The shorter-wavelength features of the geoid cannot arise from deep density variations, but are caused mainly by the topography of lithospheric blocks that, for features larger than a few hundred kilometers, are isostatically compensated by flow in the asthenosphere. Therefore Fig. 13.1(b) can be used to infer the stress state in the lithosphere. Comparison with the stress map (Fig. 11.11) shows that there is a correlation with the stress state. The correlation between long-wavelength features of the geoid and lithospheric stress allows the long-wavelength geoid features to be identified with the lower mantle. The asthenosphere is not completely inviscid, but for features hundreds of kilometres in extent it adjusts to isostatic equilibrium in thousands or tens of thousands of years. On this time scale it decouples the lithosphere from the deeper mantle, where density anomalies are presumed to persist for tens or hundreds of millions of years.

13.5 Stress regimes of continents and ocean floors

We now apply the arguments of Section 13.4 to oceanic and continental lithosphere. Figure 13.2 is a schematic cross-section of the lithosphere for the Africa plate from the Indian Ocean ridge to the mid-Atlantic ridge. Applying Eq. (13.25) to this situation,

$$\Delta\sigma = \frac{g}{z_\mathrm{L}}\left\{\int_0^e \rho_\mathrm{c}z\mathrm{d}z + \int_e^{R+e}(\rho_\mathrm{c} - \rho_\mathrm{w})z\mathrm{d}z + \int_{R+e}^{R+e+t}(\rho_\mathrm{c} - \rho_\mathrm{c})z\mathrm{d}z + \int_{R+e+t}^{z_\mathrm{C}}(\rho_\mathrm{c} - \rho_\mathrm{a})z\mathrm{d}z + \int_{z_\mathrm{C}}^{z_\mathrm{L}}(\rho_\mathrm{m} - \rho_\mathrm{a})z\mathrm{d}z\right\}, \tag{13.28}$$

where depths and densities are as shown in Fig. 13.4. By using average densities for each of the layers the above integrals can be written

$$\Delta\sigma = \frac{g}{2z_\mathrm{L}}\sum_1^5 (z_{i+1}^2 - z_i^2)\Delta\rho_i, \tag{13.29}$$

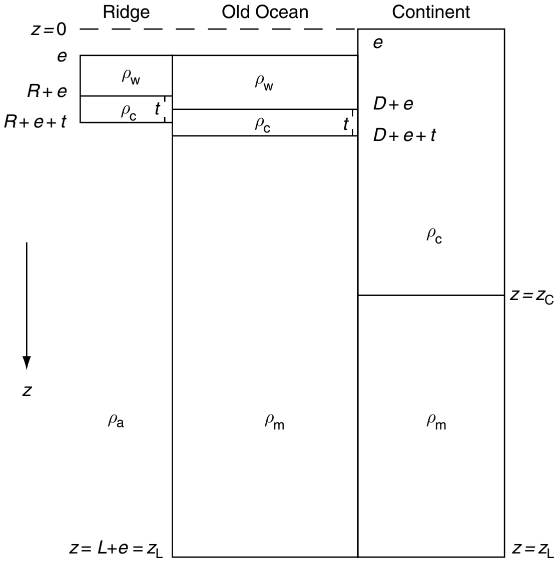


FIGURE 13.4 Layer thicknesses and densities for ocean ridges, deep ocean and continents used in calculating lithospheric stresses.

Table 13.1 Numerical values of layer thicknesses and densities in Fig. 13.4, used in calculating lithospheric stresses. *R*, *D* and *L* are depths below sea level of the ridge, deep ocean floor and base of the lithosphere, respectively

R	2500 m	ρ_w	1020 kg m^{-3}
D	6000 m	ρ_c	2800 kg m^{-3}
L	100 km	ρ_a	3300 kg m^{-3}
		ρ_m	3392 kg m^{-3}

Table 13.2 Stress difference, stress ratio and crustal thickness for different lithospheric elevations. Crustal thickness, z_C, is estimated from isostatic balance for various assumed values of surface elevation, *e*

Elevation e (m)	$\Delta\sigma = \overline{\sigma}_{xx} - \overline{\sigma}_{zz}$	σ_{zz}/σ_{xx}	z_C (km)
−6000	39.6 MPa	0.974	7.00
0	19.75 MPa	0.987	31.04
1000	10.33 MPa	0.993	36.78
1963	0	1.0	42.30
2000	−0.43 MPa	1.0003	42.51
3000	−12.51 MPa	1.0084	48.24
4000	−25.88 MPa	1.0176	53.97
5000	−40.49 MPa	1.0278	59.71
6000	−56.3 MPa	1.0390	65.44

with z_i and $\Delta\rho_i$ being the depths and density differences in the terms in Eq. (13.28).

For the comparison of the continent with the ridge the isostatic constraint on the densities and layer thicknesses is

$$\rho_C z_C + \rho_m(z_L - z_C) = \rho_w R + \rho_c t + \rho_a(z_L - R - e - t). \tag{13.30}$$

Similar equations apply to the comparison of old ocean and ridge. Table 13.1 gives numerical values for the application of this equation. Using these values in Eq. (13.28) and the corresponding equation for the old ocean-ridge comparison, we obtain the stresses and crustal thicknesses listed in Table 13.2 for various surface elevations, relative to sea level (including deep ocean). This is a simplified model and the lithosphere is very variable, so these results must be taken as a guide and not a definitive statement on crustal stresses, but they suffice to show what to expect. We have assumed a flat Earth, with Cartesian force balance, a perfect fluid at the centres of ridges and a perfectly fluid asthenosphere. The development as summarized in Table 13.2 refers to ocean floors and continents adjacent to spreading ridges with no intervening subduction zones. Tensions are observed in areas of high elevation, above about 2 km. Regions of lower elevation (with negative $\Delta\sigma$ in Table 13.2), are in compression. In this situation reverse faulting is expected. This is consistent with the mechanisms of earthquakes in old ocean areas (Wiens and Stein, 1985). More elevated continental areas are subjected to extensional stresses that cause normal faulting.

The model results in Table 13.2 assume that the depth of compensation, below which low asthenospheric viscosity ensures effectively hydrostatic conditions, is 100 km. If we assume 60 km instead, then the transition from compressional stress to extensional stress occurs at an elevation of 1 km, instead of 2 km, and conversely, compensation at 140 km depth raises this transition to 3 km elevation. We can now consider what this means for major mountain ranges (Himalaya and Andes) and high plateaux (East Africa). In the East African rift zone we have a large area of 2 km elevation and this is consistent with the idea that higher crust falls apart under the tensional stress calculated in Table 13.2. The high Himalaya and Andes require recognition of the tectonic forces that are actively pushing them up, imposing additional compressive stresses, south–north in the Himalaya (Kong and Bird, 1996) and west–east in the Andes, causing reverse faulting in the foothills. On the crests of these ranges normal faults, that mark the collapse process, are transverse to the ranges (Yin, 2000; Yuan *et al.*, 2000; Zho *et al.*, 2001). The additional stress produces higher mountains than are allowed in the quasistatic situation assumed in Table 13.2. The high Himalaya (Coblentz *et al.*, 1998) and Andes (Lamb and Davis, 2003) appear to arise from special cases of plate convergence in which subduction is inhibited over sections of these ranges by increased friction from an absence of the normal lubrication by wet marine sediment.

Since the compressive stress apparent in old ocean, even as it approaches a subduction zone, is presumed to be driven by push from a distant ridge, the implication is that the asthenospheric drag on oceanic plates is not significant, and therefore that the asthenospheric viscosity is very low. This is deduced also in Sections 13.2 and 13.4. It is probably less true under continents, which tend to 'drag their feet' in the plate motions, but it is still possible to relate the features of the world stress map (Fig. 11.11) to the same forces (e.g. Richardson, 1992).

The emphasis on ridge push appears to relegate subduction zones to a secondary role in the generation of tectonic stresses, but this emphasis must be tempered by recognition that the energy of thermal convection, which drives the whole tectonic process, is derived from vertical motion, and that this is most obvious in the subduction zones. Tectonic stresses may appear to be derived from the elevations of ocean ridges, but this is due to buoyancy and cannot be isolated from the upward displacement of hot mantle material by subduction of cooled lithosphere. The ridge elevation above old ocean floor is a measure of the lithospheric contraction that gives the negative buoyancy driving subduction.

Loads on the lithosphere with dimensions up to a few hundred kilometers can be supported by elastic flexure and depart from local isostasy. Such support has been called regional isostasy in that the balancing force is spread over a larger region than that occupied by the load. For example, the Hawaiian island chain has depressed the surrounding sea floor by several kilometers. By modelling this as a load on an elastic plate floating on an inviscid asthenosphere (e.g. Watts, 2001) and fitting either topography or gravity to the model, an effective thickness of the elastic plate can be determined. Elastic thicknesses of about 30 km, no more than 30% of the seismologically estimated lithospheric thickness, are inferred, as discussed in Section 20.4.

13.6 Coulombic thrust wedges

Earthquakes in subduction zones occur at all depths to about 700 km, but most of the major ones are shallow, occurring in the upper few tens of kilometres on shallow dipping ($\sim 20^\circ$) slabs. Because of overburden pressure the mechanism of frictional sliding of deep earthquakes requires that either deviatoric stresses are extremely large or, the more likely explanation, that the effective friction is low because of increased pore fluid pressures (Eq. 11.42). The second of these is consistent with global models of stresses within plates (Bird, 1998), which favour effective friction coefficients that have low values of about 0.1 to 0.2 compared with values determined in the laboratory of 0.6–0.85 (Section 11.5). The material in the seismogenic zone above a slab undergoes brittle deformation that involves piling up of wedges of scraped-off sediments and in some cases crustal deformation and thickening in the overriding plate. In this section we examine the associated failure and stresses.

We consider the pile of sediments or crustal rock that builds up above a subducted plate using a simplified treatment of the analysis by D. Davis *et al.* (1983). The sedimentary rocks form a wedge of material, the so-called accretionary wedge, which deforms continuously as it grows. The form of the deformation includes folding of the rock or thrusting where slabs of rock push over other slabs along low-angle planes called decollement surfaces. The time averaged behaviour of the fold and thrust belts, so formed, can be treated as if the material is in a continuous state of failure, according to Coulomb's equation (Eq. 11.41). Such a material is called a Coulombic material. As the wedge develops, the slope of the wedge topography adjusts to balance the stresses within it. A bulldozer pushing a pile of sand up a hill represents an analogous process. Initially the sand deforms internally and the slope steepens, until an equilibrium taper is reached. The wedge then moves uphill and grows self-similarly as more material is added at the toe. The critical taper is the condition for which the wedge is on the verge of failure everywhere, including slip along its base. For the case of a wedge accreting on a subducting plate, the dipping plate forms the sloping interface, while the arc or over-riding plate, which provides the compressive stress, σ_{xx} in Fig. 13.5, corresponds to the bulldozer in the sand-pile analogy.

We examine the process for a sub-aerial wedge in order to relate frictional properties to topographic slope, α, and dip of the base, β. The effective friction depends on $(1-\lambda_{\mathrm{H}})\mu_{\mathrm{f}}$ where μ_{f} is the coefficient of internal friction (Eqs. 11.41 and 11.42) and λ_{H} is the ratio of water to lithostatic pressures. As an example, we apply the theory to calculate the friction on the base of the wedge formed by subduction of the Asia plate beneath Taiwan, where α and β are known and λ_{H} in the wedge is determined from pressure measurements in boreholes.

Let x be measured along the base and z at right angles (Fig. 13.5). Consider the forces on an element dx of the wedge per unit length in the y-direction. The gravitational force/length resolved in the x-direction is $-\rho gH\mathrm{d}x\sin\beta$. The frictional force/length is $-\tau_{\mathrm{b}}\mathrm{d}x$. If σ_{xx} is the normal stress acting across any face perpendicular to the x-axis (as elsewhere in this chapter, we assume the rock mechanics convention with compression positive), the net force in the x-direction is $\int_0^H \frac{\mathrm{d}\sigma_{xx}(x)}{\mathrm{d}x}\mathrm{d}z$, where H is the thickness of the wedge at x. Balance of these forces in the x-direction gives

$$\rho gH\sin\beta + \tau_b + \int_0^H \frac{\mathrm{d}\sigma_{xx}}{\mathrm{d}x}\mathrm{d}z = 0. \tag{13.31}$$

We assume that the vertical stress arises from the weight of the material. In the small angle approximation for β,

$$\sigma_{zz} = \rho g(H-z). \tag{13.32}$$

Allowing for fluid pressure by introducing the Hubbert–Rubey coefficient, as in Eqs. (11.41) and (11.42), the effective normal traction becomes

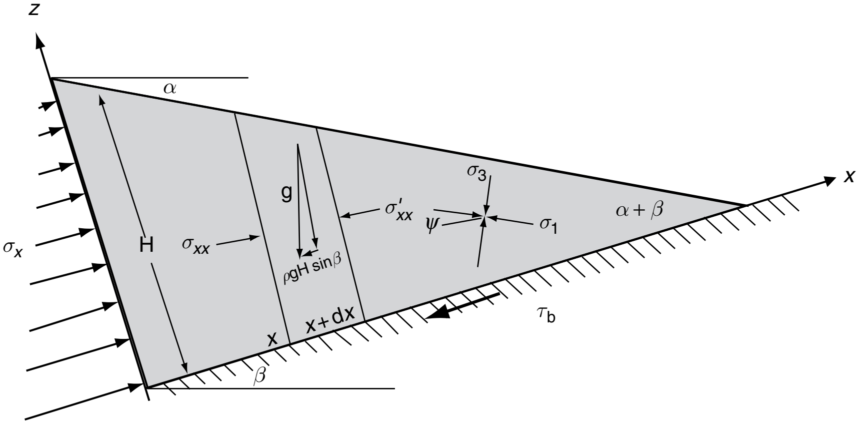


FIGURE 13.5 Geometry of an accretionary wedge. The gradient of an x-directed compressional stress in the wedge balances the x-component of gravity and friction on the base.

$$\sigma^*_{zz} = (1 - \lambda_w)\rho g(H - z). \qquad (13.33)$$

The traction at the base is

$$\tau_b = \mu_b\sigma^* = \mu_b(1 - \lambda_b)\rho gH, \qquad (13.34)$$

and failure in the wedge gives

$$\tau_w = \mu_w\sigma^* = \mu_w(1 - \lambda_w)\rho g(H - z), \qquad (13.35)$$

where μ_b is the coefficient of friction along the base and λ_w,λ_b are the Hubbert–Rubey ratios at an arbitrary level in the wedge and at the base. We treat the cohesionless case, that is, $S_0 = 0$ in Eq. (11.42). For the base of the wedge to be a non-slip surface, the frictional stress must be less than the stress required to cause internal deformation, which means that $(1 - \lambda_b)\mu_b \leq (1 - \lambda_w)\mu_w$.

Regional Coulombic failure means that any depth within the wedge σ_{xx} adjusts to be greater than σ_{zz} by an amount that generates sufficient shear to cause failure. For a critical wedge H is linear in x and so $\sigma_{xx} \propto \sigma_{zz}$. Let

$$\sigma_{xx} = A\sigma_{zz}. \qquad (13.36)$$

Using Eq. (13.32) and $dH/dx = -\tan(\alpha + \beta)$, Eq. (13.31) becomes

$$\rho gH\sin\beta + \tau_b - A\rho gH\tan(\alpha + \beta) = 0, \qquad (13.37)$$

where the proportionality constant, A, is determined from the failure criterion as follows.

The assumption that the whole wedge is at the point of failure means that, if we know A, we can use Eq. (13.36) to obtain $(\sigma_{xx} - \sigma_{zz}) = (A - 1)\sigma_{zz}$ and substitute for σ_{zz}, using Eq. (13.32), to obtain σ_{xx}. Consider the Mohr circle of Fig. 13.6(a). The maximum and minimum effective compressive stresses are σ^*_1 and σ^*_3. Let the angle between the axis of maximum compression σ^*_1 and the x-axis be ψ. At the top of the wedge σ_1 must be parallel to the topography, so $\psi = \alpha + \beta$ and, for shallow dips, we make the approximation $\psi(z) =$ constant, that is, ψ is independent of depth. (The more general case applicable to steeper dips is treated by D. Davis *et al.*, 1983). Since the wedge material is taken to be at failure at every point, the failure criterion defines two planes that are at the point of failure and are oriented at angles $\pm(\pi/4 - \phi/2)$ with respect to the σ_1 axis (Fig. 13.6(b)). From triangle BCA in Fig 13.6(a),

$$\frac{1}{2}(\sigma^*_{xx} - \sigma^*_{zz}) = \frac{1}{2}(\sigma^*_1 - \sigma^*_3)\cos 2\psi, \qquad (13.38)$$

and from triangle OAC,

$$\frac{1}{2}(\sigma^*_1 - \sigma^*_3) = \frac{1}{2}(\sigma^*_{xx} + \sigma^*_{zz})\sin\phi. \qquad (13.39)$$

Combining Eqs. (13.32), (13.33), (13.38) and (13.39),

$$\frac{1}{2}(\sigma^*_{xx} - \sigma^*_{zz}) = \frac{1}{2}(\sigma_{xx} - \sigma_{zz}) = \frac{(1 - \lambda_w)\sigma_{zz}}{\csc\phi\sec 2\psi - 1}. \qquad (13.40)$$

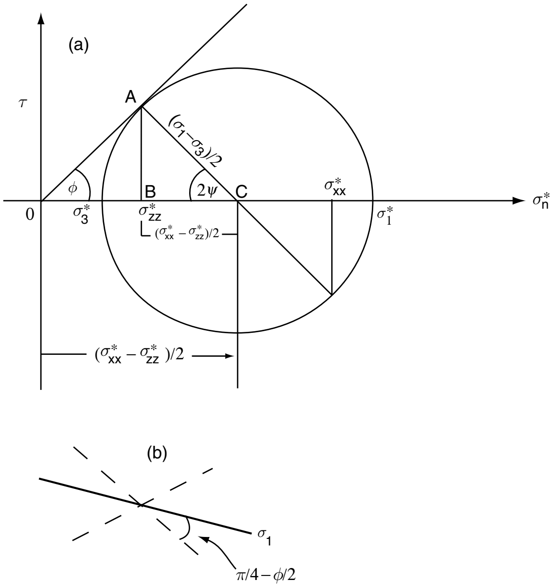


FIGURE 13.6 Mohr circles (a) within the wedge. The geometry for the maximum principal stress, and failure planes (dotted) in the wedge are shown in (b).

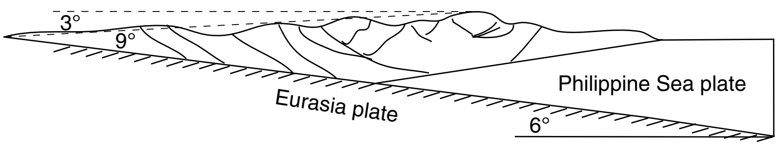


FIGURE 13.7 Application of Coulombic failure theory to the Taiwan accretionary wedge, modified after D. Davis *et al.* (1983).

Solving for A in Eq. (13.36),

$$A = \left\{1 + \frac{1 - \lambda_w}{\csc\phi \sec 2\psi - 1}\right\}. \tag{13.41}$$

then, with substitution for τ_b by Eq. (13.34), Eq. (13.37) becomes

$$\alpha + \beta = \frac{(1 - \lambda_b)\mu_b + \beta}{1 + (1 - \lambda_w)/(\csc\phi \sec(2\alpha + 2\beta) - 1)}. \tag{13.42}$$

Then

$$\lambda_b = 1 - \{(\alpha + \beta)[1 + (1 - \lambda_w)/ (\csc\phi \sec(2\alpha + 2\beta) - 1)] - \beta\}/\mu_b. \tag{13.43}$$

Thus, if we assume a laboratory value of $\mu_f = \mu_w = \mu_b = 0.85$, the friction on the base can be calculated from measurements of α, β and λ_w.

An example of application of this formula is shown in Fig. 13.7 for the Taiwan wedge. The Asian plate is subducting beneath west Taiwan forming a mountain belt that is described as a Coulombic wedge. The surface slope, α, has a mean value of about 3.0°. Offshore seismic reflection profiling has determined the slope β of the decollement surface at the toe of the wedge to be 6°, and it is presumed to remain constant landwards. Fluid pressures measured in deep boreholes indicate $\lambda_w = 0.7$. If fluid pressure were hydrostatic, the value would be about 0.4. Higher values mean that the fluid is over-pressured; it has been trapped in the sediments and compressed as the formation has been buried. We assume the laboratory values for friction coefficients (the Byerlee law, Section 11.5), $\mu_b = \mu_w = 0.85$. This leaves one parameter to be determined in Eq. (13.42), λ_b. Substituting these values in Eq. (13.43) gives $\lambda_b = 0.85$ for the base, which satisfies the condition $(1 - \lambda_b)\mu_b \leq (1 - \lambda_w)\mu_w$, that is the wedge material has a larger effective friction coefficient (Eq. 11.38), $\mu_w^* = (1 - \lambda_w)\mu_w = 0.25$, than the base $\mu_w^* = (1 - \lambda_w)\mu_w = 0.13$. This is consistent with the global value of 0.1 to 0.2, mentioned in the opening paragraph of this section.

In another example, the importance of variable friction is seen in the greater topographic slope of the central high Andes compared with the southern and northern low Andes. Because of the dry climate, the trench in front of the high Andes is almost devoid of sediment, whereas that to the north and south has abundant sediments. The main source of water at the wedge base and in the wedge itself is thought to come from dehydration of subducted sediments. Greater effective friction on the dry slab can account for the larger wedge and elevation of the central Andes and its cumulative effect over time accounts for the indentation of South America (Problem 13.7).

The critical wedge theory (using a more advanced treatment than described here) has been applied to fold and thrust belts and accretionary wedges worldwide, including the Himalaya, the fold and thrust belts on the landward side of the Andes, western North America (Horton, 1999) and the Zagros mountains of Iran (Bird, 1978). It is used to examine the interrelated effects of pore pressure, failure, sedimentation, erosion and thrusting. Near the toe of the wedge, the base slips steadily, but at greater depths, in the seismogenic zone, slip occurs quasi-periodically in large earthquakes. Stress that causes sliding along the base is given by Eq. (13.34), which, for the Taiwan case, becomes $\tau_b = 0.13\,\rho g H$. At a depth of 20 km this is 76 MPa, that is an average deviatoric stress of 38 MPa over this depth range. This value is comparable to the internal deviatoric stress in old ocean derived from ridge push, 37 MPa (Table 13.2), but is more than two orders of magnitude higher than the asthenospheric shear stresses (0.1 MPa) calculated in Section 13.2.

The magnitude of deviatoric stress, τ, in slabs is limited by the convective energy allowed by the thermodynamic arguments in Section 22.4. The analysis in this section demonstrates the importance of pore pressure for frictional sliding at depth. For Taiwan we find $\lambda_b = \lambda_H = 0.85$ compared with 0.7 in the wedge itself. At greater depths λ_H must increase to near 1.0 if frictional siding is the mechanism of deep earthquakes, so that $\tau \geq \mu_f(1 - \lambda_H)\sigma_n$. Otherwise overburden pressure would preclude earthquakes at depths greater than a few tens of kilometres.

14

Kinematics of the earthquake process

14.1 Preamble

Elucidation of the details of the tectonic pattern is one of the incentives for seismicity studies. Another is earthquake prediction. In the 1960s and early 1970s there was a widespread, but not universal, expectation that a decade or so of intensive research would yield a methodology for predicting the times, places and magnitudes of earthquakes. The task was underestimated because the physical mechanism of earthquakes is not as simple as was supposed. In spite of the lack of success we now have a somewhat clearer perception of the earthquake process, although rather little encouragement to believe that detailed prediction is possible. Nevertheless, research with this aim continues, so we can expect further improvement in our understanding of earthquake mechanisms.

The underlying driving power for earthquakes is derived from thermal convection of the mantle, the subject of a thermodynamic analysis in Chapter 22. We can compare the energy released by earthquakes with the energy that is shown thermodynamically to be available. Except for a few shallow shocks, for which fault displacements are directly observed, earthquake energies are estimated from the radiated elastic waves. Energy is directly related to magnitude and, given the numbers of earthquakes as a function of magnitude, we can integrate over all earthquakes to estimate the total energy release. Typically, the efficiency for conversion of static elastic energy to seismic waves is estimated as 6%. Although there is substantial uncertainty in this number, it suffices for an order of magnitude estimate of the fraction of total convective energy that is released as earthquakes, ~3%. Most of the mantle deformation occurs aseismically. Implications for rheology and the pattern of mantle convection are considered in Sections 12.5 and 13.2.

Most earthquakes occur at depths too great to allow access to the source zones with boreholes or tunnelling. The earthquake process is generally inferred from radiated seismic waves, but direct observation of surface displacements in some shallow shocks give convincing confirmation of the mechanism. To a first approximation an earthquake is a dislocation (Fig. 10.5), with displacement of the rock on opposite sides of a fault that grows from a starting point, termed the hypocentre. The spectrum of radiated waves is characteristic of fault dimensions and can be used to estimate them. The lowest frequencies, (longest wavelengths) are used to estimate earthquake strength or moment. Seismic moment is the product of slip, area of a fault and elastic modulus. Detailed seismic analysis of waves from very large earthquakes, such as the 2004 Sumatra–Andaman event, reveals that, at any instant, movement is restricted to a limited area of the fault, a slipping patch, that travels from one end of the fault to the other, breaking material in front of it and leaving behind broken material that eventually ‘heals’. Waves ‘pile up’ in the forward direction and are spread out in the backward direction. This directivity, apparent as an asymmetry in the radiation pattern, is a Doppler effect, and is used to determine the direction and speed of fault propagation. Modern seismic arrays reveal fine details, indicating that large dislocations can be regarded as superpositions of many sub-events.

Plate tectonic reconstructions and space geodesy have been used to calculate slip across tectonic boundaries. Comparison with the slip evident in earthquakes allows an estimate of the fraction of the plate motion in a given tectonic region that is accounted for by earthquakes and not by aseismic creep. Because very large earthquakes are rare, summing moments from historic earthquakes is inadequate. Most of the slip takes place in the largest events. Section 14.7 describes a method of estimating the effects of the rare large events by a statistical analysis of the smaller events.

Large thrust and normal earthquakes in the sea floor cause vertical motions that generate tsunamis. Only earthquakes with fault dimensions hundreds of kilometres in extent are effective. They cause tsunamis, with wavelengths comparable to the fault dimensions, that propagate in oceans, $\sim$5 km deep, as shallow water waves, meaning that they have wavelengths much greater than the water depth. The waves travel across the ocean at a speed more than an order of magnitude faster than normal wind driven waves but less than the speeds of seismic waves by a similar factor. Seismic signals are used as a basis of tsunami warning systems, supplemented by sea floor pressure observations.

14.2 Earthquakes as dislocations

Earthquakes result from rapid fault slip but only rarely do the faults break the surface. Shallow transcurrent (strike-slip) faults, across which horizontal displacements occur at the surface, provide an exception. The San Andreas fault in California is a notable example. Thus, the seismicity of California has played a vital role in recognizing earthquakes as fault movements. An example is shown in Fig. 14.1. The fact that repeated movements occur in the same direction, and accumulate as large fault offsets, is illustrated in Fig. 14.2, making it obvious that earthquakes are local increments in the pattern of tectonic motion.

Impressive as horizontal movements on transcurrent faults often are, it is important to recognize that they have a secondary role in the tectonic scheme. The energy is ultimately derived from convection, for which the essential, primary mass movements are vertical. The world's major thrust faults (see Fig. 14.10(b)),

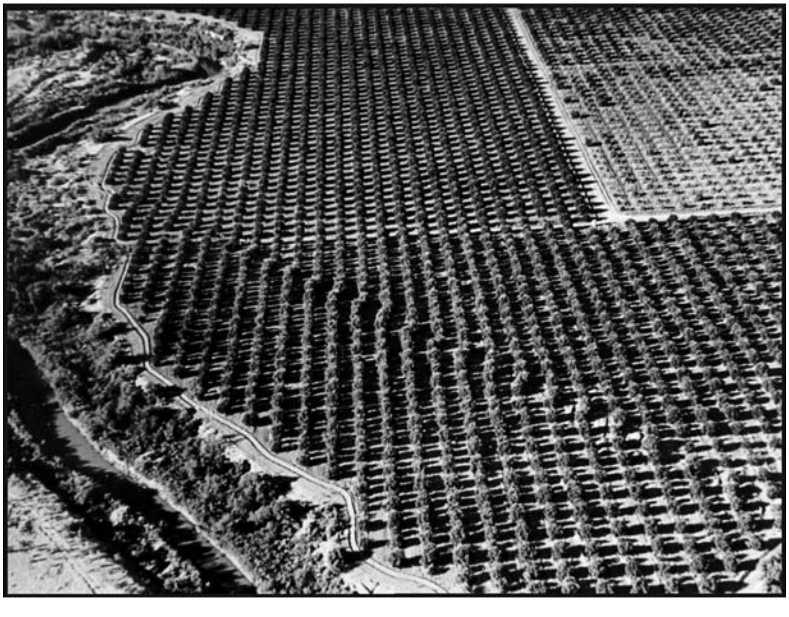


FIGURE 14.1 Fault displacement that occurred though an orange grove during an earthquake in Imperial Valley, California, in September, 1950. Photograph by David Scherman, *Life Magazine* © Time Inc.

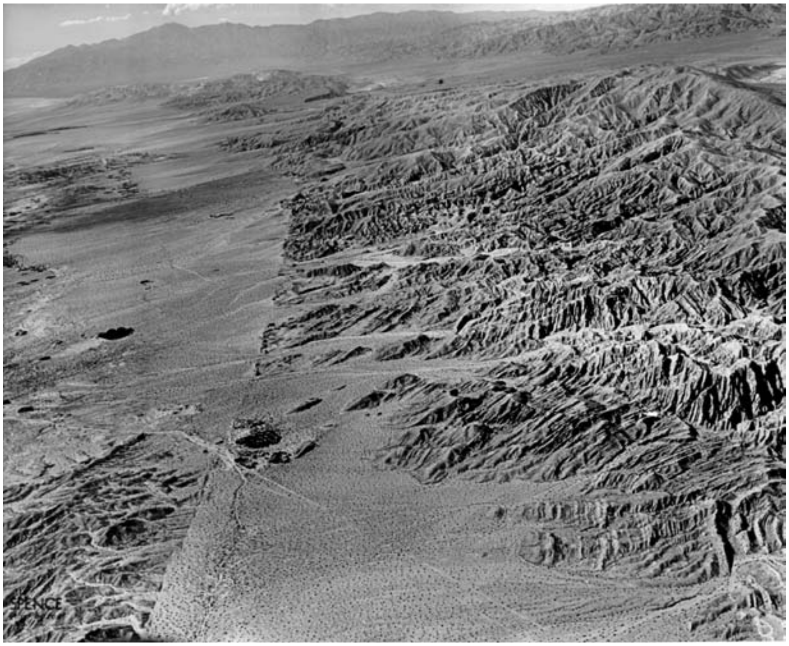


FIGURE 14.2 Mismatch of geological features across a branch of the San Andreas fault near Indio, California. Photograph by Spence Air Photos.

across which vertical slip occurs, are responsible for most of the moderate and large earthquakes, but they are generally deeper and are not directly observed. Decades of observations of fault movements so strongly emphasized horizontal displacements that they probably delayed by many years recognition of the underlying tectonic mechanism.

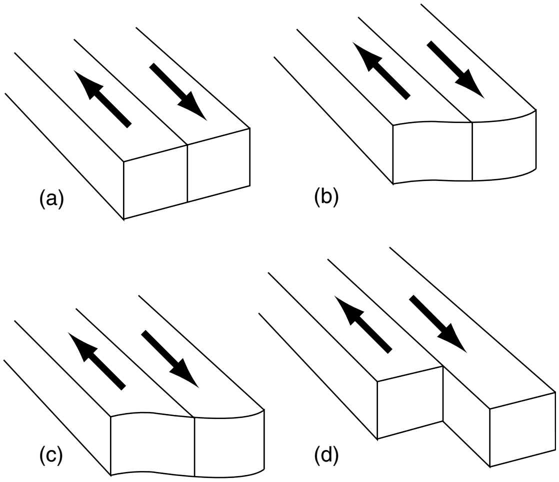


FIGURE 14.3 The elastic rebound theory of an earthquake, as advanced by Reid (1910) to explain the San Francisco 1906 event. Regional shearing movement gradually builds up elastic strain from state (a) to state (c), at which it is suddenly released across the fault, producing the displacement in (d).

The San Francisco earthquake of 1906 has a special place in the history of seismicity studies. A geodetic survey of the area had been completed shortly before the earthquake and was repeated after it, providing detailed, quantitative documentation of the displacements. Reid (1910) used the data to support what he called the elastic rebound theory of earthquakes. The theory was not entirely new, but Reid's evidence made it convincing and it became central to ideas about the earthquake mechanism. The elastic rebound theory is represented diagrammatically in Fig. 14.3. It envisages a progressive regional movement, causing elastic deformation of the ground until a breaking point is reached, when displacement across a fault releases the elastic strain. The concept of a progressive regional movement causing build-up of elastic strain, which is abruptly released in occasional events, falls naturally into place with current ideas about plate tectonics. Thus the essential validity of the elastic rebound theory is not subject to doubt. However, the concept of a breaking point is simplistic. We still have no clear idea of what triggers earthquakes or what limits the spread of a fault break once it starts. This is a difficulty that limits progress in earthquake prediction.

Fault displacements, and the associated elastic strains and stresses, are described by equations that are collectively termed dislocation theory. Some of the jargon is adopted directly from solid-state physics in which there is an extensive literature on dislocations. They occur as defects in crystals and control their strengths and mechanisms of deformation. There are two basic kinds, illustrated in Fig. 10.5. Consider a solid body of convenient shape, make a half-cut in it and displace the material on opposite faces of the cut. The axis of the dislocation is the edge of the cut and the relative displacement of the faces is the displacement or Burgers vector. If this is parallel to the axis, as in Fig. 10.5(a), then the result is a screw dislocation. If the displacement vector is perpendicular to the axis (Fig. 10.5(b)), it produces an edge dislocation, and in general we have mixed dislocations, with both types of displacement.

To a useful approximation, the San Francisco 1906 earthquake can be modelled as a screw dislocation. Another intensively studied earthquake, Alaska 1964, resembles a pair of edge dislocations. These simple models represent faults that are much longer in one dimension than the other. The simple dislocations envisaged in Fig. 10.5 are described by two-dimensional equations, that is, they assume uniformity along the axis or an effectively infinite length. More general equations are needed for faults that are nearer to equidimensional (see Section 14.3). Another generalization that is needed for realistic application to earthquakes is a grading of slip to zero at fault boundaries. For faults that break the surface it is the gradient of the slip that must be zero at the surface.

Ground displacements accompanying the San Francisco 1906 earthquake extended for hundreds of kilometres along the San Andreas fault, but were concentrated in a band about 15 km either side of it. This means that the break was shallow and quite well represented by a screw dislocation, although the slip was variable along the length of the fault. Figure 14.4(a) gives the geometry for a simple screw dislocation that models this situation, with its axis parallel to the surface at depth D. Initially we calculate the strain for a dislocation in an infinite medium and then take account of the free surface by introducing a hypothetical image dislocation above the surface. For the infinite medium the circle of radius r about the dislocation axis, which is shown as a

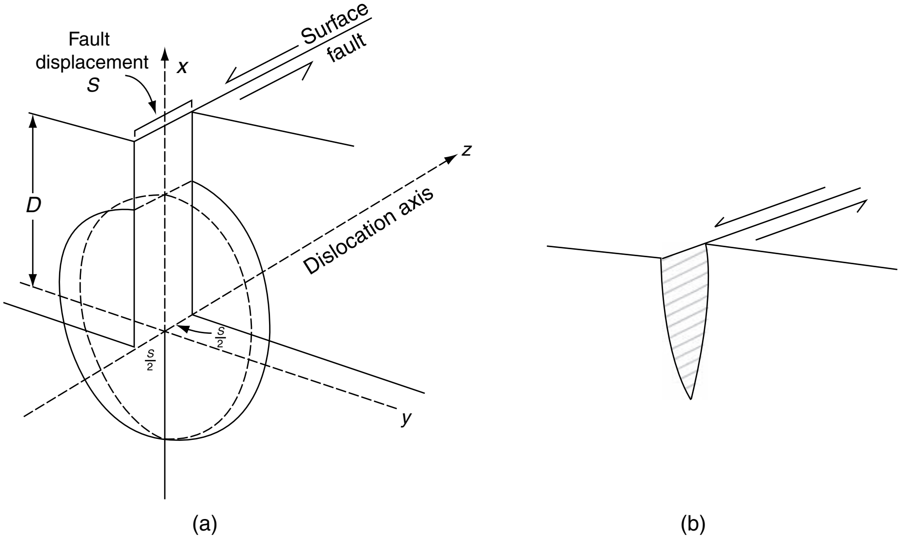


FIGURE 14.4 (a) Geometry of a screw dislocation, the mathematical model of a transcurrent fault. Displacement, S, is uniform to depth D and zero below that, twisting the circle (broken line) into a helical form. (b) A more realistic variation of displacement with depth.

broken line for the unfaulted medium, is deformed to the helical curve (solid line). The shear strain, ε, is uniform around the circumference of this circle and is therefore given by

$$\varepsilon = b/2\pi r, \tag{14.1}$$

where b is the displacement (Burger's vector). We are primarily interested in the component ε_{yz} of this strain across a vertical plane, $y = \text{constant}$, parallel to the fault plane ($y = 0$),

$$\varepsilon_{yz} = \frac{b}{2\pi r}\frac{x}{r} = \frac{b}{2\pi}\frac{x}{(x^2 + y^{2)}}. \tag{14.2}$$

This gives the strain component in an infinite medium, due to a single dislocation of displacement, $S = b$.

We may now consider the effect of a free surface at $x = D$. Since there are no stresses across the surface, the shear strain across it, ε_{xz}, is zero everywhere on the plane. Mathematically, this situation can be achieved by introducing to the infinite medium a hypothetical mirror image dislocation at $x = 2D$. The effect of the two dislocations together would be to give slip S for $0 < x < 2D$ but zero outside this range. The second dislocation does not cancel ε_{yz}, which becomes

$$\varepsilon_{yz} = \frac{b}{2\pi r}\left(\frac{x}{x^2 + y^2} + \frac{2D - x}{(2D - x)^2 + y^2}\right). \tag{14.3}$$

In the free surface ($x = D$), where observations are normally made,

$$\varepsilon_D = \varepsilon_{yz}(D) = \frac{bD}{\pi(D^2 + y^2)}. \tag{14.4}$$

Thus displacements of surface points at arbitrary distance, y, from the fault are obtained by integrating the strain, ε_D, with respect to y, noting the discontinuity in displacement at $y = 0$,

$$\text{displacement} = \int_{-\infty}^{y} \varepsilon_D \mathrm{d}y = \frac{\mathrm{b}}{2}\left(1 - \frac{2}{\pi}\tan^{-1}\frac{y}{D}\right). \tag{14.5}$$

Equation (14.5) gives the surface displacement (in the z direction) due to a simple dislocation with a discontinuity of fault displacement on the dislocation axis at depth D ($x = 0$). There are singularities in stress and strain on the axis and the model can be improved by grading the fault slip to zero as $x \to 0$ (but keeping the variation of slip with depth zero as $x \to D$) as in Fig. 14.4(b). The effect of these refinements is too slight to affect the fit of the geodetic observations (Fig. 14.5) made after the 1906 earthquake, and the curve shown is for a simple dislocation. Other unmodelled effects include variability of displacement along the fault and heterogeneity of the medium. Recently much finer observations using radar interferometry from satellites have provided detailed displacement maps that allow heterogeneity to be identified (Peltzer *et al.*, 1999). A cross-section of ground displacements resulting from the 1996 Manyi

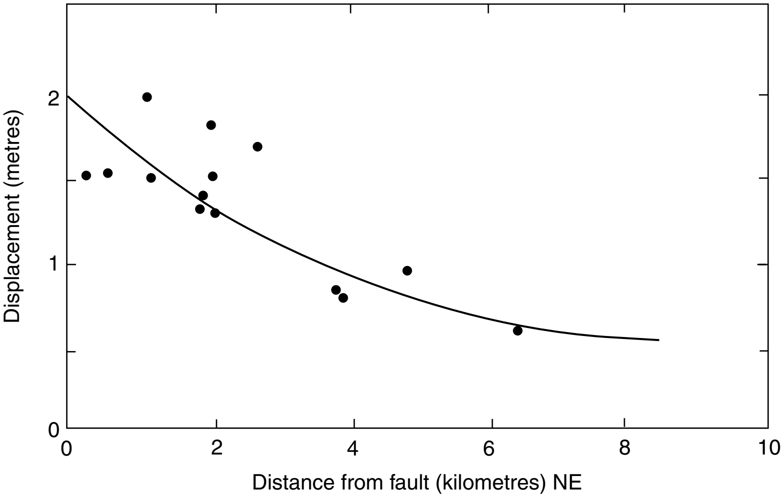


FIGURE 14.5 Horizontal displacement of surface points on the NE side of the San Andreas fault that occurred during the San Francisco 1906 earthquake. The curve represents Eq. (14.5) with $S = 4$ m and $D = 3.4$ km. To allow for grading of slip, as in Fig. 14.4(b) a somewhat greater total fault depth must be favoured.

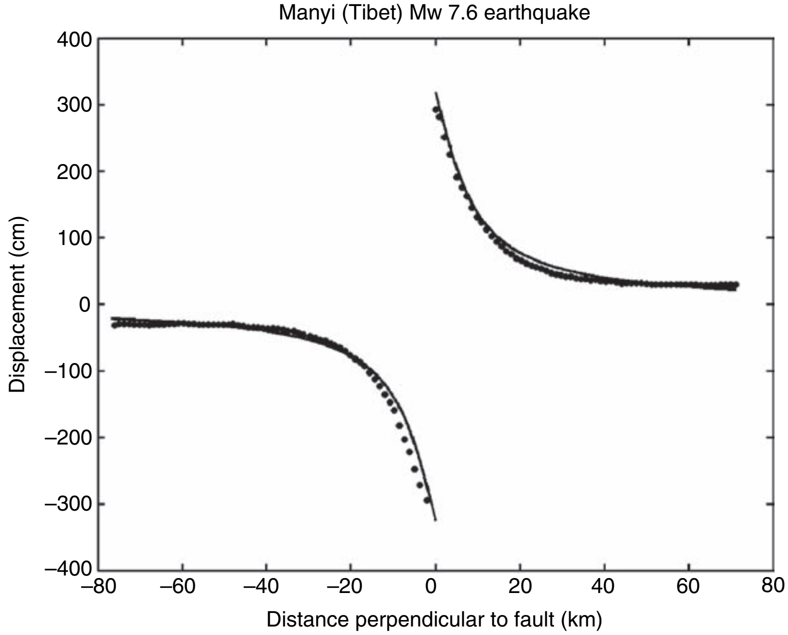


FIGURE 14.6 Displacements determined from InSar (interferometric synthetic aperture radar) for the 1996 Manyi, Tibet, earthquake, from data by Peltzer *et al.* (1999). A fit to Eq. (14.5), with $S = 6.84$ m, and $D = 7.9$ km, of data provided by Gilles Peltzer.

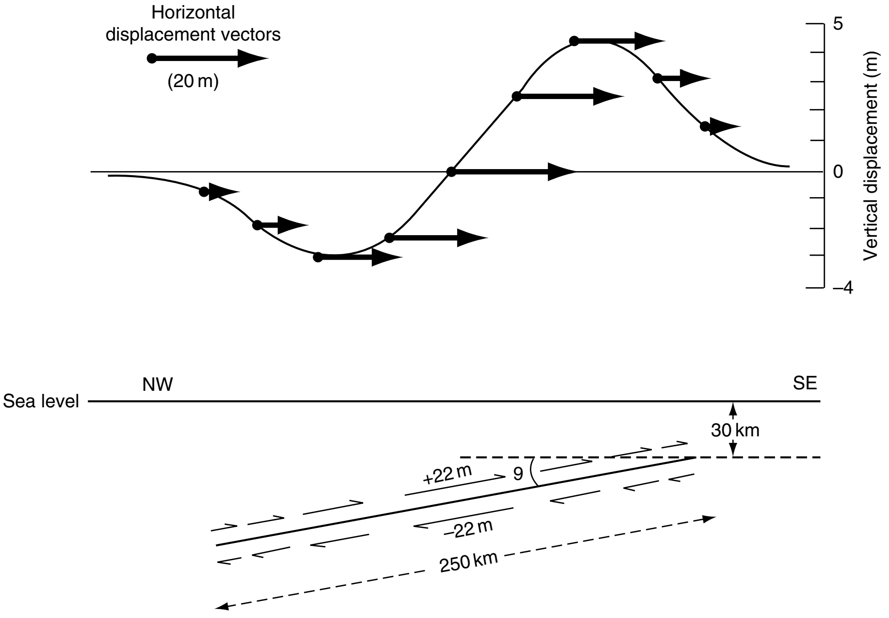


FIGURE 14.7 A model of fault movement during the Alaska 1964 earthquake, matched to observed displacements of the surface reported by Plafker (1965). The model assumes that the fault did not break the surface, but if it did so the break would have been at sea.

strike-slip earthquake in Tibet (Fig. 14.6) show a remarkable fit to predictions of Eq. (14.5), but asymmetry of other cross-sections led Peltzer *et al.* to conclude that elastic response about the fault was variable.

A model of the Alaska 1964 earthquake is shown as the lower part of Fig. 14.7, with corresponding surface displacements, in the top half of the figure, matched to the available observations. The Alaska 1964 earthquake cannot be modelled as simply as the San Francisco 1906 earthquake. It is represented by a pair of compound edge dislocations with graded displacements. This is a two-dimensional model. The fault length, roughly parallel to the Alaska coast, was about 800 km, four times the width and sufficient to make the two-dimensional model satisfactory over the central section of the fault. Equations for edge dislocations, not presented here, are not as simple as those for screw dislocations, partly because the free surface cannot be accounted for by a single, simple image. Equations for displacements due to arbitrarily oriented faults are given by Mansinha and Smylie (1971) and Okada (1985).

We can estimate the maximum strains, close to the faults, for the two earthquakes considered and use the elasticity of the Earth at appropriate depths to obtain the corresponding stresses. For San Francisco we may take the movement on each side of the fault to be 2 m and distribute this displacement over the depth of the fault. The simple dislocation model gives the depth as about 3.5 km, but a detailed analysis, using a more realistic compound dislocation model, gives (5.0 ± 1.5) km. It is puzzling that the movement on a fault break hundreds of kilometres in length should be so shallow, but using the latter figure, the strain is 4×10^{-4} and taking the crustal rigidity as 30 GPa the drop in shear stress across the fault was 12 MPa (120 bar). This is an upper bound, as slip was variable along the fault and a high value is assumed. On other fault sections the stress drop may not have exceeded 2 MPa. For the Alaska 1964 shock we distribute the 22 m maximum displacement over the half-width of the fault, 125 km, to obtain a strain of 1.8×10^{-4}. At the greater average depth of this shock the shear modulus is about 55 GPa, giving a stress drop of 9.9 MPa (99 bar). Acknowledging the uncertainties in these calculations, we can guess that a stress drop of order 10 MPa (100 bar) is typical of large, shallow shocks. Estimates in the range 1 to 10 MPa are derived from studies of the spectra of seismic waves (Section 14.5). These stresses are much less than the breaking stress of previously unfractured rock. Most earthquakes follow lines of pre-existing weakness.

14.3 Generalized seismic moment

Even for earthquakes on faults that break the surface, the pattern of surface displacements is normally very irregular and incompletely recorded and there are no direct measurements of displacement at depth. Moreover, we are interested in the sequence of events during an earthquake, including details such as the point of initiation (hypocentre) and the speed of fault propagation. To obtain this information we must decipher the records of the radiated elastic waves. For many years earthquakes have been located by timing wave arrivals at widely distributed observatories and Chapter 17 discusses the use of seismic wave arrival times to infer the Earth's internal structure. This chapter is concerned with the information that can be obtained about earthquakes themselves. Although they are complicated phenomena and are certainly not all the same, some clear patterns have emerged.

The screw and edge dislocations discussed in the previous section are simplified two-dimensional models of earthquake sources. In three dimensions, dislocation patches are used, as developed in this section. The measure of earthquake strength is seismic moment, which takes into account the elasticity of the medium, the amount of displacement, and the area over which the displacement occurs. The scalar seismic moment M_0 is given by

$$M_0 = \int \mu b \mathrm{d}S, \qquad (14.6)$$

where μ is rigidity modulus of the faulted medium and b is the distribution of slip across a fault of area S. In three dimensions seismic moment becomes a tensor, considered below.

Calculation of the radiation field from an earthquake fault is a complicated mixed-boundary problem, in which stress (friction) and displacement conditions must be simultaneously satisfied on both faces of the fault. The problem is simplified by replacing the fault by combinations of point forces in a continuous elastic medium that can be adjusted to be equivalent to motions across a discontinuous surface. Such point forces have relatively simple analytical solutions and so can be superimposed to represent any internal dislocation of the medium.

A point force in an infinite medium can be thought of as a force that is exerted at a point, a finger that is inserted and pushes in a particular direction. The elastic material resists this push and the material moves until the elastic restoring force balances the disturbance. The material is compressed in front of the force and expands behind it. The force may be dynamic, i.e. varying in time, or static.

For a point force, F_k, applied at the origin, the equilibrium equations (Eq. 11.43) become

$$\frac{\partial \sigma_{ij}}{\partial x_j} + F_k \delta_{ik} = 0, \tag{14.7}$$

where repeated indices are summed. The solution to this equation is the Green's function (Landau and Lifshitz, 1975),

$$G_{ik} = \frac{1}{4\pi\mu}\left[\frac{\delta_{ik}}{r} - \frac{1}{4(1-\nu)}\frac{\partial^2 r}{\partial x_i \partial x_k}\right], \tag{14.8}$$

where G_{ik} is the displacement in the i-direction from a unit point force in the k-direction. The point force on its own has limited application in geophysics. A special case of a point force applied to the surface of an elastic half space, the so-called Boussinesq problem, is used in modelling surface loads. One example is the load imposed on the surface by an impounded lake. Gough and Gough (1970) used a distribution of half-space point forces to describe the displacements and stresses associated with the filling of Lake Kariba in Zimbabwe. Comparison with the induced seismicity and deformation measured by levelling confirmed that elastic constants determined for the upper crust from seismology also describe static loading.

By differentiating Eq. (14.8) we obtain equations describing dislocations in three dimensions that are much more general than the equations of the previous section (Eshelby, 1973). Opening of a dike or tension crack is modelled by a third type of dislocation in which the Burgers (displacement) vector is normal to the crack surface. A cavity inflation is modelled in the far field, by three surfaces at right angles with displacement normal to each, whereas a collapse corresponds to displacement in the opposite sense. Each of these cases can be modelled by combinations of derivatives of the point force equations (Fig. 14.8).

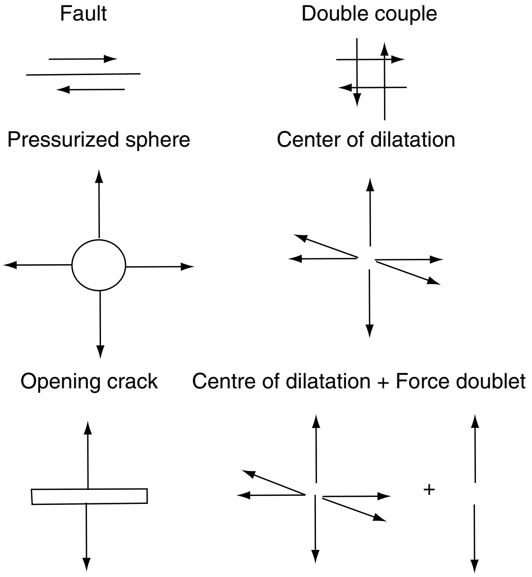


FIGURE 14.8 Point force representations of elementary dislocations.

For each case of dislocation sliding, opening or volume collapse, the force combination must have no net force and no net moment, because the end result has neither linear nor rotational acceleration. For an earthquake dislocation model, slip is tangential to the fault plane and in opposite directions on opposite sides. We model the slip on the two sides in terms of oppositely directed point forces, which together form a couple. However, a couple has a net torque, so its sudden appearance at the time of an earthquake would develop an accelerating angular momentum. To avoid this, a second couple must be added with an equal and opposite torque and forces at right angles to the first. The result is the double-couple model of earthquakes.

To form the individual couples we use the same mathematical procedure as in electrostatics to form a dipole. Suppose that the two surfaces of the displaced crack are a distance δz apart, which in the limit we will decrease to zero. Let the normal to the crack area, $\mathbf{S}$, be parallel to the z-axis, and the horizontal displacement, $\mathbf{b}$, in the x-direction. The strain is $b/\delta z$. Then the force is $\mu(b/\delta z)S$. The displacement generated by the upper and lower forces is

$$\begin{aligned} u_i(\text{couple}) &= \text{force}\left\{G_{ix}\left(r+\frac{\delta z}{2}\right) - G_{ix}\left(r-\frac{\delta z}{2}\right)\right\} \\ &= \frac{\mu b S}{\delta z}\left\{G_{ix}\left(r+\frac{\delta z}{2}\right) - G_{ix}\left(r-\frac{\delta z}{2}\right)\right\} \\ &= \mu b S\frac{\partial G_{ix}}{\partial z}, \end{aligned} \tag{14.9}$$

where μbS is seismic moment (Eq. 14.6), and G is the Green's function as in Eq. (14.8). Adding a compensating couple at right angles, of the same strength but opposite torque, we have

$$u_i(\text{double couple}) = \mu bS\left(\frac{\partial G_{ix}}{\partial z} + \frac{\partial G_{iz}}{\partial x}\right). \tag{14.10}$$

Here the force couple is made up of two equal and opposite forces displaced normal to their line of action. Seismologists use a similar construction to model the opening of a crack. The forces in this case are displaced in the direction of opening or parallel to their line of action to form a force doublet. Three doublets added together at right angles make a centre of dilatation.

$$u_i(\text{centre of dilatation}) = \lambda bS\left(\frac{\partial G_{ix}}{\partial x} + \frac{\partial G_{iy}}{\partial y} + \frac{\partial G_{iz}}{\partial z}\right). \tag{14.11}$$

A centre of dilatation is used to model a spherical region that expands. Combinations of unequal doublets can be used to model the expansion of ellipsoidal regions (Davis, 1986). A much-used application of the centre of dilatation solution is the Mogi model of stressing of a volcano by a magma chamber (Anderson, 1936; Mogi, 1958). However, in that case extra image terms are added to the equations to cancel tractions on the surface of the Earth.

Opening of a crack is modelled by combining a centre of dilatation and a doublet oriented in the direction of opening. The full expression is

$$\begin{aligned} u_i(\text{tension crack}) = bS\Bigg(\lambda&\left(\frac{\partial G_{ix}}{\partial x} + \frac{\partial G_{iy}}{\partial y} + \frac{\partial G_{iz}}{\partial z}\right) \\ &+2\mu\frac{\partial G_{iz}}{\partial z}\Bigg), \end{aligned} \tag{14.12}$$

where λ is the Lamé parameter (see Section 10.2). In general **b** varies over the dislocation surface and to model finite faults the solutions are integrated and may include time variation, as in Section 14.5.

Equations (14.10) to (14.12) may be generalized to incorporate both shear and dilatation,

$$u_i = M_{pq}\frac{\partial G_{ip}}{\partial x_q}, \tag{14.13}$$

where M_{pq} is the seismic moment tensor, defined in connection with Eq. (14.6), and given by

$$M_{pq} = \lambda\delta_{pq}b_pS_q + \mu(b_pS_q + b_qS_p), \tag{14.14}$$

where b and S are components of the vectors **b** and **S** and $\frac{\partial G_{ip}}{\partial x_q}$ is the displacement in the i-direction from a point force in the p-direction, differentiated in the q-direction to form a couple or doublet.

Equation (14.13) gives the displacement in the i-direction for an average Burgers vector, b, dislocating a surface, S, in terms of derivatives of the equations for displacements by elementary point forces, that is, in terms of couples and doublets. The shear crack lying in the x–y (1–2) plane has a moment tensor given by

$$M_{pq} = \begin{pmatrix} 0 & 0 & \mu b_1 S \\ 0 & 0 & 0 \\ \mu b_1 S & 0 & 0 \end{pmatrix}, \tag{14.15}$$

and the tension crack in the y–z (2–3) plane,

$$M_{pq} = \begin{pmatrix} (\lambda+2\mu)b_1S & 0 & 0 \\ 0 & \lambda b_1 S & 0 \\ 0 & 0 & \lambda b_1 S \end{pmatrix}. \tag{14.16}$$

Equation (14.15) demonstrates a fundamental non-uniqueness in inverting measured displacements to determine which plane has slipped in an earthquake. Suppose that we invert Eq. (14.13) to determine M_{pq} and obtain values corresponding to Eq. (14.15). Because of the symmetry of the tensor we cannot distinguish between slip on the z plane in the x direction and slip on the x plane in the z direction, with $M_{xz} = M_{zx}$. If slip occurs on the x-plane then the z-plane is referred to as the auxiliary plane. This is a consequence of using the point representation of an earthquake, as reconsidered in the following section. If we invert displacement data from a finite source we can obtain the extent and orientation of the fault plane, i.e. M_{pq} (x,y,z) as discussed in Section 14.5. The generalized moment tensor of an earthquake is obtained by inverting Eq. (14.13) where the dynamic Green's function (Aki and Richards, 2002) is used. One can rotate the tensor, as we did for stress in Chapter 11, to obtain the fault plane and the auxiliary plane. The moment can then be represented as a fault plane solution for which the plane and the slip across it are specified. For a shear dislocation, such as an earthquake, the trace of the moment tensor is zero, that is, there are no opening or closing sources (cracks).

14.4 First motion studies

The direction of the initial movement in a seismic wave, at an observing station remote from an earthquake, depends in a systematic way on the orientation of the fault and direction of slip relative to the station. In an infinite, uniform medium the pattern of compressional (P-wave) first motions would be simple. The medium would be divisible into four quadrants, with one opposite pair experiencing initial compression (motion away from the source) and the other pair initial dilation (motion towards the source). If we consider the double couple the heads of the force arrows will form compressions and the tails extensions. The boundaries of the quadrants are the fault plane and the auxiliary plane perpendicular to it, as in Fig. 14.9(a). The radiation pattern is shown in Fig. 14.9(b), as viewed in a plane perpendicular to both the nodal (fault and auxiliary)

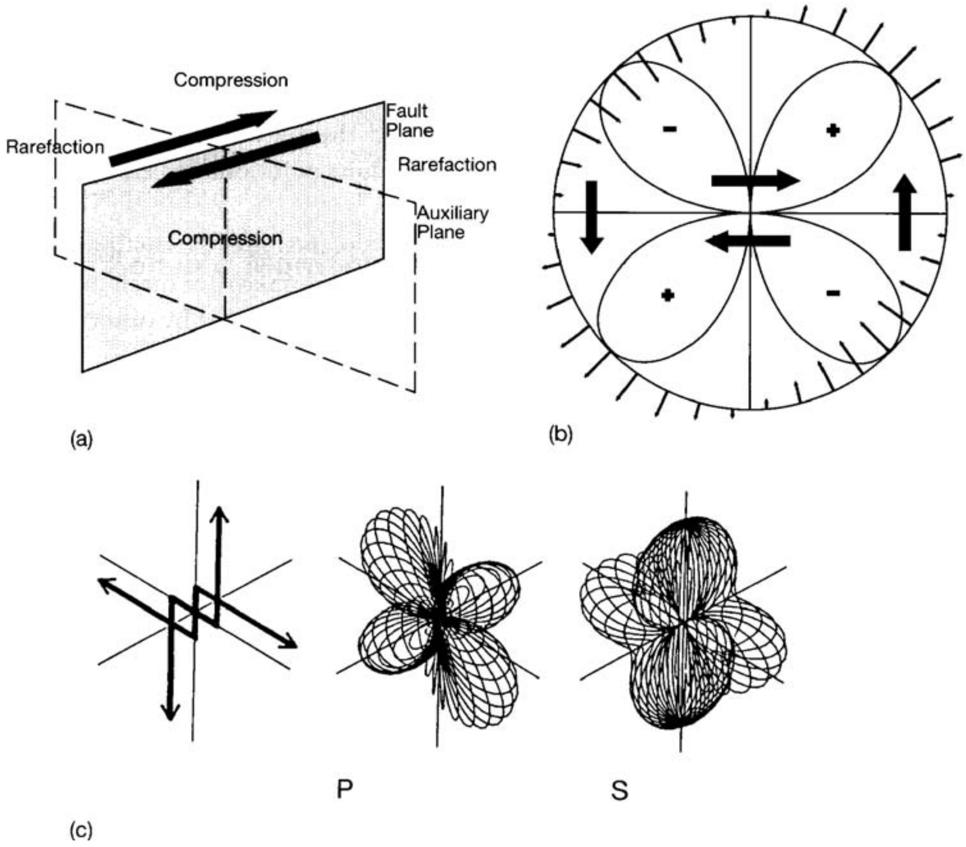


FIGURE 14.9 (a) Separation of compressions and rarefactions of the first motions of seismic P waves by the fault plane and an auxiliary plane perpendicular to it. (b) P-wave radiation pattern viewed in a plane perpendicular to the fault and auxiliary planes. At angle ϕ to this plane the amplitude is reduced by the factor $\cos\phi$. (c) A three-dimensional view of the radiation pattern from a double couple point source with the orientation shown (from Kennett, 1983, p. 90).

planes. In directions at ϕ to this, third plane the amplitude of motion is diminished by the factor $\cos\phi$ and there is no P-wave radiation in either of the nodal planes. Figure 14.9(c) gives a three-dimensional perspective.

In the Earth, seismic waves are refracted by the velocity variations and so there are corresponding refractions of the nodal planes. By allowing for these refractions, first motions observed at widely scattered stations are used to deduce the nodal planes of well-observed earthquakes and the directions of slip. The results are known as fault plane solutions, or focal mechanisms. The coverage of the Earth with reliable seismic stations suffices to allow focal mechanisms to be determined routinely for all large earthquakes and many smaller ones. They demonstrate that the movements are all contributions to the grand pattern of global tectonics and they are used to study local details of the pattern.

Without further information, the two nodal planes are indistinguishable. As we concluded from Eq. (14.15), this is a consequence of the double couple nature of earthquakes. The pictorial representation of focal mechanisms is the projection onto a horizontal plane of black (compression) and white (rarefaction) quadrants for downgoing waves, that is P waves into a hemisphere below the source. Figure 14.10 gives three simple examples. While this representation incorporates the ambiguity in the nodal planes, the directions of compressive and extensional stress are not ambiguous and the pattern of movement is often evident in a plot showing several fault plane solutions for earthquakes in a limited area.

There are several methods by which fault planes may be distinguished from auxiliary planes. Extended faults of large earthquakes are often outlined by foci of aftershocks. The direction of motion may be indicated by observations of surface displacements or by asymmetries in the wave patterns. If an earthquake is unilateral in the sense that the fault break propagates strongly in one direction from its initiation, but hardly at all in the opposite direction, then the amplitude of the first pulse of the P-wave train is greater in the direction of propagation but its duration is shorter. The radiation pattern departs from the symmetry of Fig. 14.9(b) and the ambiguity of a conventional fault plane solution is then resolved. We examine this phenomenon, called ‘directivity’, in the next section.

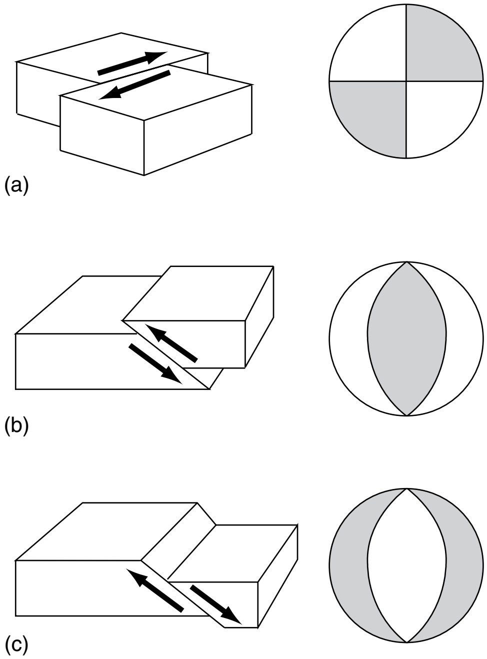


FIGURE 14.10 Focal mechanisms for earthquakes on (a) a transcurrent fault with no vertical movement, (b) a thrust fault, of the kind occurring in subduction zones, and (c) a normal fault. The ‘beachball’ representation of first motions shows the quadrantal pattern as seen from below the faults, with shading for initial compressions.

First motion studies have generally emphasized P waves, but S waves give complementary information. In this case the motion is perpendicular to the radius from a source, and has both θ and ϕ components, where θ is the angle in the plane of Fig. 14.9(b) and ϕ is the angle with respect to this plane. Using unit vectors $\hat{r}, \hat{\theta}$ and $\hat{\phi}$, as in Aki and Richards (2002), the P (radial) and S (tangential) first motions are proportional to

$$R_r = \sin 2\theta \cos\phi \hat{r} \quad \text{for P waves,} \tag{14.17}$$

$$\left.\begin{array}{l} R_\theta = \cos 2\theta \cos\phi\hat{\theta} \\ R_\phi = -\cos\theta \sin\phi\hat{\phi} \end{array}\right\} \quad \text{for S waves.} \tag{14.18}$$

14.5 Rupture models and the spectra of seismic waves

In this section we integrate the dynamic Green's function to show some relationships used to interpret seismograms, in particular the estimation of seismic moments from the low-frequency spectra of seismic waves. A point source, that is, one with infinitesimal dimensions and rupture time, would give rise to a white spectrum. Finiteness of the rise time of slip and the fault dimensions of a real source cause the white spectrum ideal to be truncated at high frequencies, but the low frequency range of the spectrum is white. This requires the frequency range considered to be very low, relative to the slip duration, and to correspond to wavelengths much greater than the fault dimension, so that the source is indistinguishable from a point.

We examine a rectangular fault model with rupture propagating at a fixed velocity, less than the S-wave speed. The rupture moves from one end of the fault to the other, with rupture occurring on a slipping patch, with a rupturing edge and a healing edge a fixed distance apart. The waves in the direction of rupture tend to have higher frequencies than those measured in the opposite direction because the source of radiation is moving in the direction of wave propagation, causing the energy to 'pile up' and give a shorter, more intense pulse. In the opposite direction the source is receding and so the waves become drawn out as the rupture front recedes. This is a seismological Doppler effect.

Equation (14.8) gives the static Green's functions that can be used in Eq. (14.13). Integration over a two-dimensional strike-slip fault gives Eq. (14.5). The Green's function for a dynamic point force is quite complicated, with near, intermediate and far field terms, but is simplified by considering the far field only. In a medium with P and S wave speeds V_{P} and V_{S}, the far field displacement caused by a fault patch is given by

$$\begin{pmatrix} u_r \\ u_\theta \\ u_\varphi \end{pmatrix} = \frac{1}{4\pi\rho r V_{\mathrm{P}}^3}\dot{M}_0(t - r/V_{\mathrm{P}})\begin{pmatrix} R_r \\ 0 \\ 0 \end{pmatrix} + \frac{1}{4\pi\rho r V_{\mathrm{S}}^3}\dot{M}_0(t - r/V_{\mathrm{S}})\begin{pmatrix} 0 \\ R_\theta \\ R_\varphi \end{pmatrix}, \qquad (14.19)$$

where $\dot{M}_0$ is the rate of change of the moment tensor associated with the patch. Equation (14.19) has the general form

$$u = \frac{1}{4\pi\rho V^3}R^{\mathrm{PS}}\frac{1}{r}\dot{M}_0(t - r/V), \qquad (14.20)$$

where V is V_{P} or V_{S} and R^{PS} is the P or S wave radiation pattern (Eq. (14.17) or (14.18)). We take the Fourier transform of Eq. (14.20),

$$u(\omega) = \frac{1}{4\pi\rho V^3}R^{\mathrm{PS}}\frac{1}{r}\int_{-\infty}^{\infty}\dot{M}_0(t - r/V)\exp(\mathrm{i}\omega t)\mathrm{d}t. \qquad (14.21)$$

Let $\tau = t - r/V$. Then

$$u(\omega) = \frac{1}{4\pi\rho V^3}R^{\mathrm{PS}}\frac{1}{r}\exp(\mathrm{i}\omega r/V)\int_{-\infty}^{\infty}\dot{M}_0(\tau)\exp(\mathrm{i}\omega\tau)\mathrm{d}\tau. \qquad (14.22)$$

We are interested in the asymptotic value of Eq. (14.22) as $\omega \to 0$. The integral term becomes $\int \mathrm{d}M_0 = M_0$. Then

$$u(\omega \to 0) = \frac{M_0}{4\pi\rho V^3}R^{\mathrm{PS}}\frac{1}{r}\exp(\mathrm{i}\omega r/V). \qquad (14.23)$$

This illustrates the important result that the displacement spectrum of far-field P and S seismograms at $\omega \to 0$ can be used to obtain the total moment. Equation (14.22) shows that, if the moment rate is a delta function, the spectrum is flat (white) for all frequencies. A more realistic case is a boxcar function of duration T, as in Fig 14.11(a), between $-T/2 \leq \tau \leq T/2$. If the slip is D over area A, $\dot{M} = \mu AD/T$ so that Eq. (14.22) becomes

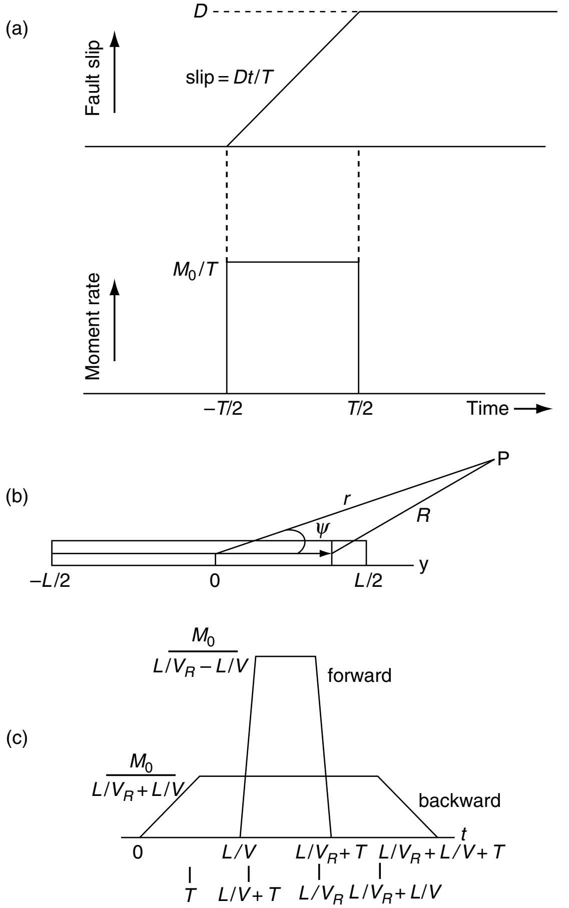


FIGURE 14.11 A fault model for calculation of the radiation pattern from an earthquake. (a) Variation of moment and moment rate with time. (b) Finite fault for the Haskell (1969) calculation. (c) Resulting seismic pulses.

$$
\begin{aligned}
u(\omega) &= \frac{1}{4\pi\rho V^3} R^{\mathrm{PS}} \frac{1}{r} \exp(\mathrm{i}\omega r/V) \int_{-T/2}^{T/2} \dot{M}_0 \exp(\mathrm{i}\omega\tau)\mathrm{d}\tau \\
&= \frac{M_0}{4\pi\rho V^3} R^{\mathrm{PS}} \frac{1}{r} \exp(\mathrm{i}\omega r/V) \sin(\omega T/2)/(\omega T/2) \\
&= \frac{M_0}{4\pi\rho V^3} R^{\mathrm{PS}} \frac{1}{r} \exp(\mathrm{i}\omega r/V) \mathrm{sinc}(\omega T/2). \qquad (14.24)
\end{aligned}
$$

The finite rise time gives a $1/\omega$ displacement spectrum at high frequencies.

To model displacements from a propagating break of a finite fault we integrate Eq. (14.20). Consider a long thin fault of width W and length L (Fig. 14.11(b)), for which rupture starts at one end and progresses to the other end at constant speed, V_{R}. This is the rupture velocity, which, for many earthquakes is found to be 0.7 to 0.9 times the S-wave speed. At any given location, once the slip starts, it rises to its final value D, which could be 1 m, in time T, and then remains constant. T is referred to as the rise time and is generally a few seconds. Thus the slip velocity, that is, the speed of the material in the fault face, is metres per second, compared with the rupture velocity, which is several kilometres per second. We can think of this as a wave, with particle motion much slower than the advance of the wave. This is the simplest finite fault model, referred to as the 'Haskell model', recognizing its originator, Norman Haskell, who also solved the isostatic rebound problem (Chapter 9). The Haskell model has five parameters, fault dimensions of length L, width W, slip D, rise time T and rupture velocity V_{R}. It has an active slipping patch that starts slipping at $-L/2$ at $t=0$ (Fig 14.11(b)). It builds up to a constant size $W \times (T \times V_{\mathrm{R}})$ and sweeps from one end to the other at the rupture velocity, rather like dominos falling in a line, with the ones falling at any instant representing the patch.

Consider an element of the moving patch with dimensions $W\mathrm{d}y$ at y. Then $\mathrm{d}\dot{M}_0 = \mu\dot{D}W\mathrm{d}y$. The radiation from this element arrives at a distant point, P (Fig. 11.4(b)), at time t that is the sum of the rupture time, $(L/2+y)/V_{\mathrm{R}}$, and travel time, R/V. We integrate over the whole fault, taking into account the fact that $\dot{D}$ is non-zero only during the period of slip of the element. This can be represented by writing $\dot{D}$, as a function of time, as $\dot{D}[t-(L/2+y)/V_{\mathrm{R}}-R/V]$, over the duration of patch slip. Then

$$
u = \frac{\mu W}{4\pi\rho V^3} R^{PS} \int_{-L/2}^{L/2} \frac{1}{R}\dot{D}[t-(L/2+y)/V_{\mathrm{R}}-R/V]\mathrm{d}y. \qquad (14.25)
$$

In the far field, $r \gg L \gg W$, $R = r - y\cos\psi$, and for the purpose of the $1/R$ term in Eq. (14.25), $r \approx R$ and $1/R$ can be taken outside the integral. Equation (14.25) becomes

$$u = \frac{\mu W}{4\pi\rho V^3} R^{\mathrm{PS}} \frac{1}{r} \int_{-L/2}^{L/2} \dot{D}\ [t - r/V - (L/2 + y)/V_{\mathrm{R}} + y\cos\psi/V]\mathrm{d}y = \frac{1}{4\pi\rho V^3} R^{\mathrm{PS}} \frac{1}{r} P(t), \qquad (14.26)$$

where the seismic pulse shape $P(t)$ is a characteristic of the source and the other terms refer to propagation of the waves. Recalling that $\dot{D} = D/T$ and $M_0 = \mu WLD$, we have

$$P(t) = \frac{M_0}{LT} \int_{-L/2}^{L/2} I[t - r/V - (L/2 + y)/V_{\mathrm{R}} + y\cos\psi/V]\mathrm{d}y, \qquad (14.27)$$

where $I[t] = 1$ for $0 \leq t \leq T$. Integration of (14.27) gives a pulse that has the shape of a trapezoid (Fig. 14.11(c)), with high amplitude, short duration and rise time in the direction of rupture and lower amplitude, longer rise time and duration in the opposite direction (Problem 14.6). The area of each trapezoid is equal to M_0 for all ψ. This is the source function and the observed pulses are modulated by propagation effects, including the radiation pattern (Eqs. (14.17) and (14.18)). An example of this directivity is seen in the pulses observed at different azimuths from 1995 Colima-Jalisco, Mexico $M_{\mathrm{W}}8$ earthquake (Courboulex *et al.*, 1997) in Fig. 14.12. From the sharpness of west-bound pulses and the longer duration of the southbound pulses we see that this event propagated in a northerly direction.

In order to investigate the frequency dependence of Eq. (14.27) we take the Fourier transform, with a change of variables $t = \tau - r/V - (L/2 + y)/V_{\mathrm{R}} + y\cos\psi/V$ and $X = (1/V_{\mathrm{R}} - \cos\psi/V)$

$$P(\omega) = M_0 \exp(\mathrm{i}\omega[r/V + T/2 + L/2/V_{\mathrm{R}}]) \frac{1}{LT} \int_{-T/2}^{T/2} \exp(\mathrm{i}\omega\tau)\mathrm{d}\tau \int_{-L/2}^{L/2} \exp(\mathrm{i}\omega Xy)\mathrm{d}y, \qquad (14.28)$$

where $\mathrm{d}t = \mathrm{d}\tau$, because the other terms in the variable substitution are constants. This gives

$$P(\omega) = M_0 \exp(\mathrm{i}\omega[r/V + T/2 + L/2/V_{\mathrm{R}}]) \times \mathrm{sinc}(\omega T/2)\mathrm{sinc}(\omega XL/2). \qquad (14.29)$$

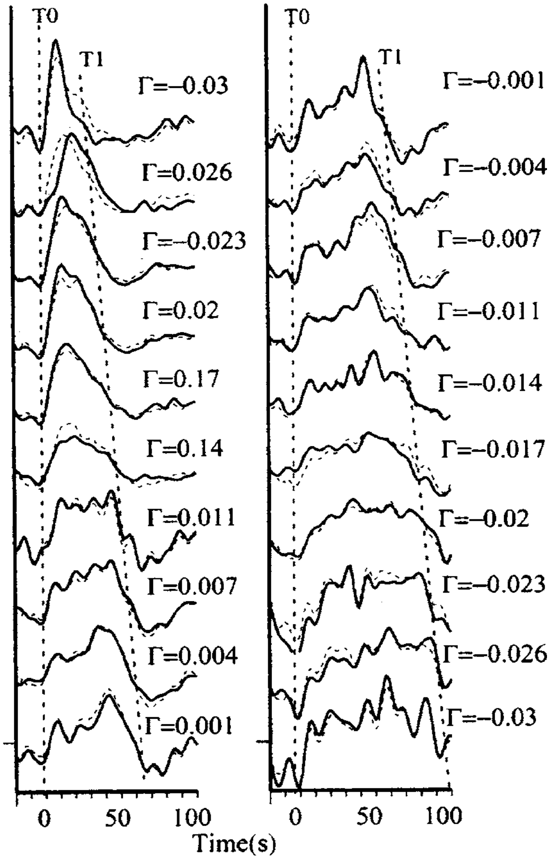


FIGURE 14.12 Source time functions at different azimuths, found by inversion of surface waves from the 1995 Colima-Jalisco (Mexico) earthquake (after Courboulex *et al.*, 1997). *T*0 and *T*1 are estimated start and end times. The pulse broadening can be used to estimate the direction of rupture, N 79° W, average rupture velocity, 2.8 km/s, and fault length, $L = 150$ km.

We now have a spectral amplitude that is a product of two sinc ($\sin x/x$) functions, resulting from the finite rise time and finite propagation time, compared with the single sinc function arising from the finite rise time in Eq. (14.24). The exponential term is a phase delay and does not affect amplitudes. The zero-frequency asymptote of Eq. (14.29) is the moment, while at high frequencies P decays as $1/\omega^2$. We see from Eq. (14.29) and Fig. 14.11 that the amplitude of the pulse may be written as the convolution of two boxcar functions.

$$P(t) = M_0 \int_{-\infty}^{\infty} B_1(t - \tau)B_2(\tau)\mathrm{d}\tau, \qquad (14.30)$$

where

$$B_1(t) = 1/\mathrm{T}, \quad 0 \le \mathrm{t} \le \mathrm{T},$$

and

$$B_2(t) = 1/(LX), \quad 0 \le \mathrm{t} \le LX. \tag{14.31}$$

The first is associated with the rise time and the second with the propagation. In the frequency domain the convolution of two boxcars is the product of two sinc functions. Since at high frequencies the amplitude of a sinc function decays as $1/\omega$, the spectrum of the finite fault decays as $1/\omega^2$. The $\cos\psi$ term in X gives the 'forward' pulse (in the direction of fault propagation) a greater high frequency content than the 'backward' pulse. At zero frequency the product becomes unity, which is why the low frequency asymptote of seismic wave amplitude yields the moment.

As mentioned, if an earthquake is viewed as an instantaneous point source, its spectrum is white. A finite rise time introduces a $1/\omega$ decay and, as we see in the above example, a finite spatial extent adds another $1/\omega$ factor. Aki and Richards (2002) discuss more general cases with two finite fault dimensions. The extra dimension introduces another sinc function and therefore another $1/\omega$ term to give a $1/\omega^3$ spectrum. However, in practice the $1/\omega^2$ spectrum gives a better fit to most events, presumably because propagation time across the smaller fault dimension is not significant.

Figure 14.13 summarizes the basic features of seismic spectra. The curves show the displacement spectra for earthquakes of different magnitudes, magnitude being a measure of earthquake size, discussed in Section 14.6. Each curve follows the equation

$$u(\omega) = \frac{u(0)}{1 + (\omega/\omega_0)^2}, \tag{14.32}$$

where $u(\omega)$ is the spectral 'amplitude' at frequency $(\omega/2\pi)$, according to the model of Brune (1970), and is similar to the double sinc function of Eq. (14.29). Thus $u(\omega)$ is the square root of spectral power per unit frequency interval. $f_0 = \omega_0/2\pi$ is termed the corner frequency in the Brune model and varies systematically with moment M_0. By Eq. (14.29), M_0 is proportional to $u(0)$ and the corresponding scale is shown on the left-hand side of the figure. This figure illustrates several of the conclusions of the spectral analysis in this section.

(i) The spectra become white (flat) at sufficiently low frequencies, allowing a direct determination of M_0.

(ii) At the high frequency limit, $u(\omega)$ varies as ω^{-2}. This is the observed general form or envelope, as the high frequency spectra of earthquakes are generally irregular, being complicated by interference phenomena arising from multiple starting and stopping events. At the highest frequencies attenuation may be a problem but this does not obscure the observation that spectral amplitude increases with magnitude or moment even in the ω^{-2} range. However, large earthquakes are observed to radiate more high frequency energy than the Brune model of Fig. 14.13 would suggest. This may be attributable to rapid acceleration and retardation of slip on small patches of a fault. Fault movement proceeds jerkily as it overcomes sticking points or asperities.

(iii) The corner frequency, f_0, which marks the boundary between low and high frequency regimes, increases systematically with decreasing M_0. f_0 is directly related to the duration of seismic radiation and since, for constant speed of fault propagation, the duration is proportional to the linear dimensions of a fault, it is a direct measure of fault size. f_0 is higher for P-waves than for the slower S-waves.

Figure 14.14 shows a sequence of images of the source of the Sumatra–Andaman earthquake (Ishii *et al.* 2005) determined by back-projecting seismograms from the HiNet array in Japan. Their analysis indicates a rupture that travelled 1300 km south to north at an average speed of 2.8 km s^{-1}. Directivity effects were seen at the very longest wavelengths and also in the azimuthal variation of the duration of the high-frequency body wave seismograms. However, at the highest frequencies the relative amplitudes did not exhibit directivity effects, presumably because the constructive interference necessary for directivity requires the high-frequency

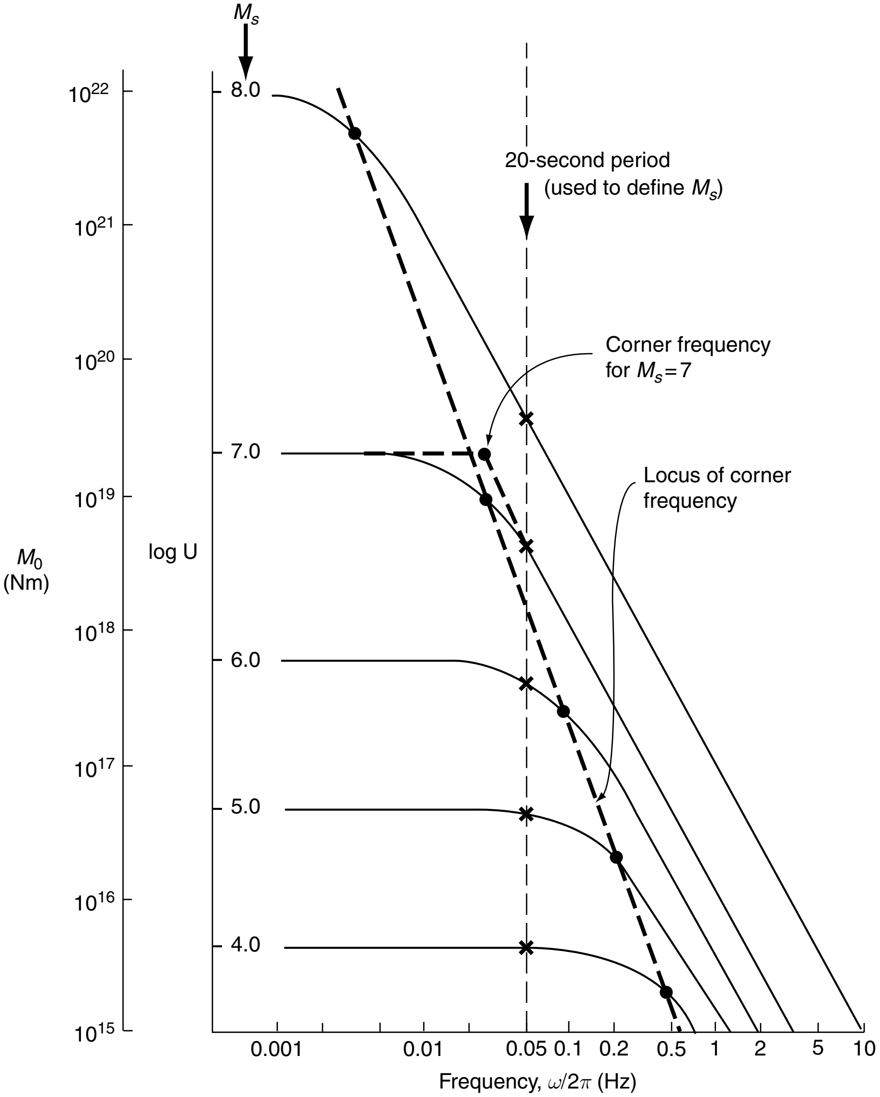


FIGURE 14.13 Idealized displacement spectra for S waves radiated by earthquakes with a range of magnitudes. The curves follow Eq. (14.32), giving the spectral amplitude $u(\omega)$, that is the square root of power per unit frequency interval (μm per $\sqrt{}$Hz), as a function of frequency, $\omega/2\pi$. u has a constant value $u(0)$ at $\omega \ll \omega_0$ but $u(\omega) \propto \omega^{-2}$ at $\omega \gg \omega_0$, where ω_0 is the corner frequency. This is the 'ω-squared model' of K. Aki (see Aki and Richards, 2002).

sources to be in phase with the propagating wavefront. While this was seen to be true for low frequencies, at higher temporal resolution the wave was erratic, growing and dying in several major pulses, as the propagating break moved northwards. A full description of broadband seismic records at high frequencies requires recognition of the stochastic nature of the sources (Chapter 15).

14.6 Earthquake magnitude and energy

Before there were instrumental recordings of earthquakes, size was reported in terms of the intensity of local ground shaking with empirical numerical scales for quantifying it. The most

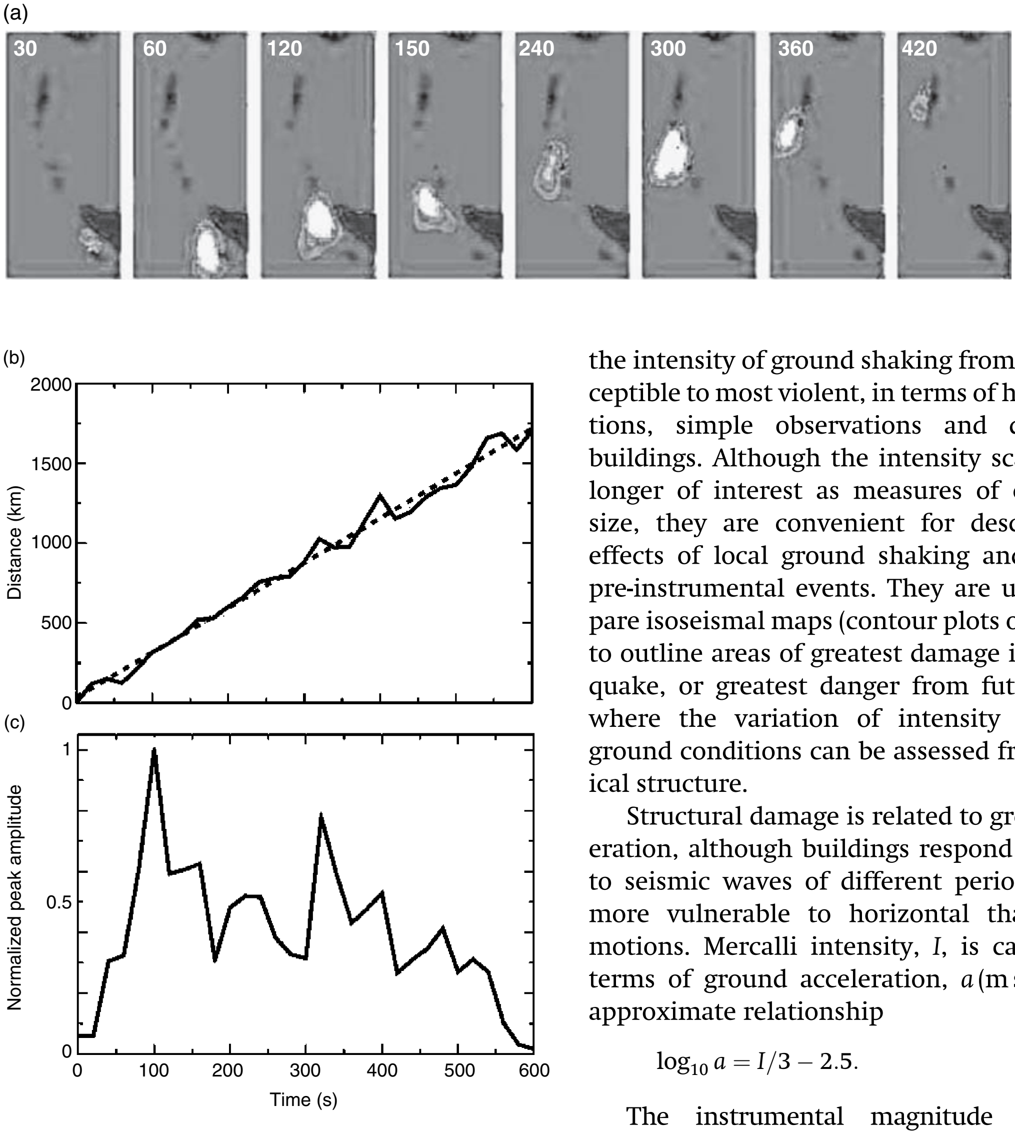


FIGURE 14.14 (a) A time sequence of fault slip for the Sumatra–Andaman 2004 earthquake, redrawn from Ishii *et al.* (2005). Light areas show active fault movement and dark shading shows land. (b) Rupture distance along the fault versus time. The dashed line is the straight-line fit to the peak locations, and gives an average rupture speed of 2.8 km s^{-1}. (c) Normalized peak amplitude as a function of time, showing two significant high-frequency energy events at ~80 s and 330 s (Ishii *et al.*, 2005).

widely used was originally proposed in 1902 by G. Mercalli and still bears his name, although there have been numerous modifications, with adaptions to local building codes. It represents the intensity of ground shaking from barely perceptible to most violent, in terms of human reactions, simple observations and damage to buildings. Although the intensity scales are no longer of interest as measures of earthquake size, they are convenient for describing the effects of local ground shaking and assessing pre-instrumental events. They are used to prepare isoseismal maps (contour plots of intensity) to outline areas of greatest damage in an earthquake, or greatest danger from future shocks where the variation of intensity with local ground conditions can be assessed from geological structure.

Structural damage is related to ground acceleration, although buildings respond differently to seismic waves of different periods and are more vulnerable to horizontal than vertical motions. Mercalli intensity, I, is calibrated in terms of ground acceleration, $a\,(\mathrm{m\,s^{-2}})$, by an approximate relationship

$$\log_{10} a = I/3 - 2.5. \tag{14.33}$$

The instrumental magnitude scale was originally developed in the 1930s, for use in California, by C. F. Richter, in collaboration with B. Gutenberg (see Richter, 1958). Californian earthquakes are shallow, and consequently generate strong surface waves, the elastic waves that are guided by the elasticity and density contrasts at and near the surface of the Earth. Surface waves generally appear on seismic records with greater amplitudes than the body waves that penetrate the deep interior. They also have longer periods, making their waveforms clearer on the records. Richter magnitudes M_L (a scale for local events) was so successful that it was extended to surface wave magnitudes, M_S, applicable to more distant earthquakes. Body-wave magnitudes, m_b, are also used. Magnitude is defined as the logarithm of the ratio of the amplitude of ground motion, A (in microns), to dominant wave period, T (seconds). With a correction term for observations at arbitrary distances Δ (degrees, subtended at the centre of the Earth) and depth h (km), the relationship is

$$M_S = \log_{10}(A/T) + f(\Delta, h), \tag{14.34}$$

to which an added correction for local seismometer site conditions may be needed. Amplitude observations over a wide range of angles from the source are needed for a reliable estimation of magnitude because an earthquake radiates differently at different orientations to the fault, but recordings at several stations are needed anyway for earthquake location so this requirement is automatically satisfied.

It is obvious that there is a general correspondence between maximum intensity, I_{max}, and M_S, although observations of high intensities in areas where ground motion is amplified by layers of soft sediment, or focussing by geological structures, can be misleading. A rough empirical relationship, with a term to account for the depth h (km) of an earthquake, is

$$M_S = 2I_{max}/3 + 1.7\log_{10} h - 1.4. \tag{14.35}$$

The rapid general adoption and success of the Richter magnitude scale is due to two features of the definition (Eq. 14.34).

(i) The logarithmic scale permits a fine subdivision over a very wide range, with representation by a simple number that never exceeds 10. (Very small earthquakes with negative magnitudes may be recorded locally.)

(ii) The ratio A/T is a measure of the strain amplitude, ϵ, in a seismic wave and, since the flux of elastic wave energy passing any point is proportional to ϵ^2, this means that magnitude is a measure of seismic wave energy.

The total energy depends also on the length, or duration, of a seismic wave-train, but by integrating complete seismic waveforms, a direct, if empirical, relationship between magnitude and energy, E (joules), is obtained:

$$\log_{10} E = 1.5M_S + 4.8. \tag{14.36}$$

Thus an increment of 1 in magnitude corresponds roughly to a 30-fold increase in energy. For some purposes it is convenient to rewrite Eq. (14.36) in an alternative form,

$$E = 6.3 \times 10^4 \exp(3.45M_S)\,\mathrm{J}. \tag{14.37}$$

With the development of wide-bandwidth seismometers, an intrinsic limitation of conventionally determined magnitudes became apparent. M_S determinations generally use Rayleigh waves (vertically polarized surface waves – see Section 15.3), recorded on instruments with periods of about 20 s. The largest earthquakes ($M_S > 8$) radiate much of their energy in waves with periods greater than this, so that the M_S scale saturates and does not give sufficient discrimination between very large shocks. The M_S magnitude scale is based on observations of surface waves with a 20 s period, and falls on different parts of the spectral curve for different magnitudes (Fig. 14.13). Note that the 20 s intersections with the curves have equal intervals. For $M_S \leq 6$, M_S is directly proportional to the logarithm of the moment, log M_0, but for $M_S \geq 7$ the 20 s period corresponds to the ω^{-2} part of the spectrum and so is an unsatisfactory measure of earthquake size. Saturation of the m_b scale, which uses P-waves, compressional body waves, often with periods of 1s but never more than 5 s, is even more serious. As we saw in Section 14.5, very wide bandwidth (or at least very low frequency) records provide a means of estimating M_0, which is a better measure of earthquake size.

The relationship between seismic moment and earthquake energy depends on the stress release mechanism, in particular the assumption that the final stress remaining when movement has ceased is equal to the frictional stress across the moving fault. The relationship is simplified by the observation that the stress release, or stress drop, is quite similar for virtually all earthquakes of magnitudes exceeding about 3, as documented for example by Kanamori and Anderson (1975). This justifies the approach of Kanamori (1977), who related M_0 to M_S for shocks in the range for which M_S is not suspect and used M_0 as the basis of a revised scale of magnitudes, M_W. This is effective for the largest shocks and coincides with M_S

for smaller ones. As in Kanamori and Brodsky (2004), the corresponding direct relationship between M_W and M_0 (in Newton-meters) is

$$M_W = (2/3)\log_{10} M_0 - 6.07. \tag{14.38}$$

The largest recorded seismic moment was for the Chile, May 1960, earthquake, $M_0 = 2.5 \times 10^{23}$ Nm, corresponding to $M_W = 9.5$ and an energy (combining Eqs. 14.37 and 14.38) of 1.2×10^{19} J.

14.7 The distribution of earthquake sizes

The number of earthquakes of magnitude M or greater per year is observed to follow a simple relationship named the Gutenberg–Richter distribution after its originators,

$$\log_{10} N = a - bM. \tag{14.39}$$

Relationships of this form, usually with $b \approx 1$, are quite generally observed in local regions as well as globally. With numerical values of the coefficients from a global data fit by Kanamori and Brodsky (2004), for $4 < M < 8$, this is

$$\log_{10} N = (8.0 \pm 0.2) - (1.00 \pm 0.03)M, \tag{14.40}$$

as plotted in Fig. 14.15. There must be limits to the magnitude range to which this applies. Although suggestions of a lower limit have appeared in the literature, studies with modern instruments, especially in deep boreholes (Abercrombie and Leary, 1993; Abercrombie and Brune, 1994), showed that Eq. (14.39) is valid down to at least $M = 0$. It is not clear that there is any lower limit short of the scale of grain boundaries in rocks. Concerning the upper limit, Eqs. (14.39) and (14.36) together show that the total energy release by earthquakes is exponentially dependent on any assumed upper magnitude cut-off. Since the total energy is finite the distribution must taper off. We consider this problem below in terms of moments rather than magnitudes.

First we consider the number–moment relationship corresponding to the Gutenberg–Richter law (Eq. 14.39). Combining Eqs. (14.38) and (14.39), we obtain

$$\log_{10} N = \log_{10} \alpha - \beta \log_{10} M_0 \tag{14.41}$$

or

$$N(\text{moment} > M_0) = \frac{\alpha}{(M_0)^{\beta}}. \tag{14.42}$$

For the purpose of these equations, M_0 must be understood as a dimensionless ratio, $(M_0/1\,\text{Nm})$ and α can be interpreted as the number of earthquakes (per year) for which $M_0 \geq 1$ Nm. With the numbers in Eqs. (14.38) and (14.40), $\alpha \approx 10^{14}$ and $\beta \approx 2/3$. Equation (14.42) is a power law

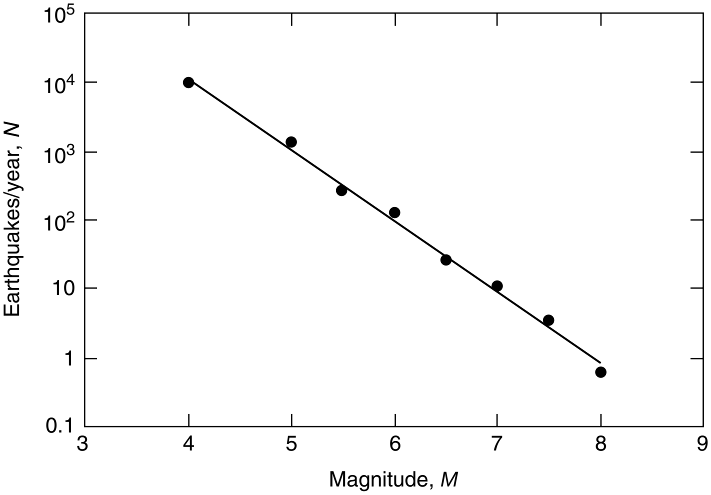


FIGURE 14.15 Magnitude–frequency relationship for all earthquakes for the period 1904 to 1980. The line has a gradient of −1, corresponding to a b-value of 1 in Eq. (14.39). This is effectively a log–log plot because M is proportional to the logarithm of wave amplitude by Eq. (14.34) or energy by Eq. (14.36). On average, approximately one earthquake with $M \geq 8$ occurs each year. Tapering of the distribution occurs at magnitudes greater than ~8.

distribution with fractal dimension β. The essential feature of fractals is that they involve no intrinsic scale, and over most of the observed magnitude range earthquakes provide a classic example. The physical processes of earthquakes appear to be independent of size (they are 'self-similar'), subject to an upper limit that we now discuss.

To calculate total moment in a given catalogue we differentiate (14.42) to obtain $\mathrm{d}N/\mathrm{d}M_0$ and then integrate $-M_0\,\mathrm{d}N$ (negative because M_0 decreases with N). With the range of M_0 unrestricted,

$$\begin{aligned} M_{0\text{total}} &= -\int_0^\infty M_0 \mathrm{d}N = \int_0^\infty \alpha\beta M_0^{-\beta} \mathrm{d}M_0 \\ &= \frac{\alpha\beta}{1-\beta}\left[M_0^{(1-\beta)}\right]_0^\infty . \end{aligned} \tag{14.43}$$

For $\beta < 1$, Eq. (14.43) is infinite. Thus, either a maximum cut-off moment (such as the Chile earthquake) must be assumed or the distribution modified to give convergence of the total moment. Kagan (1991) modified Eq. (14.42) by introducing an exponential term,

$$N(\text{moment} > M_0) = \frac{\alpha}{M_0^\beta} \mathrm{e}^{-\frac{M_0}{M_C}}, \tag{14.44}$$

giving a roll-off in the distribution at a corner moment M_C. The fit of the global data set to Eq. (14.44) is plotted in Fig. 14.16, with the assumption that a common value of $M_C = 1.2 \times 10^{22}$ Nm is applicable to the entire data set. Equation (14.44) imposes a soft upper bound that can be determined from earthquake catalogues for which statistically significant numbers of events are available to demonstrate the taper. Bird and Kagan (2004) found that corner magnitudes, m_C, corresponding to corner moments, M_C, vary between 5.9 and 9.6, depending on tectonic province, but that β does not vary much from the

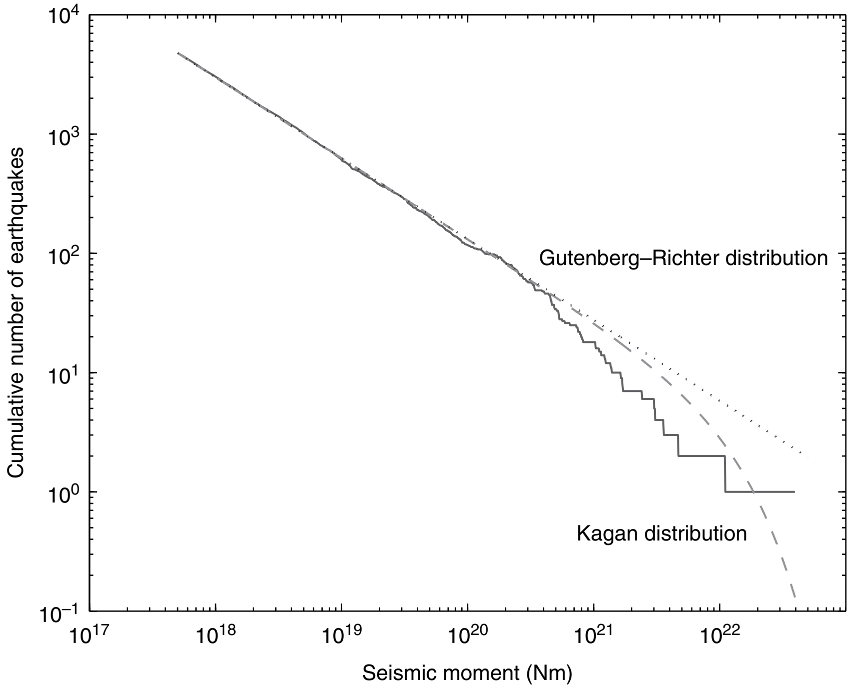


FIGURE 14.16 Moment distribution for shallow earthquakes (0–70 km) from the Harvard CMT catalogue 1977–2005, showing the evidence for a corner or taper, departing from the Gutenberg–Richter distribution at large magnitudes. The broken line is a plot of Eq. (14.44) with $M_C = 1.2 \times 10^{22}$ Nm. The break in gradient at $M_0 \approx 4 \times 10^{20}$ Nm corresponds to $M = 7.7$. The departure from linearity at this point is significant, although statistics are uncertain in the high moment range, with the small number of events. The single event at $M_0 > 4 \times 10^{22}$ Nm is Sumatra 2004. Figure provided by Yan Kagan.

Table 14.1 Analysis of the Harvard earthquake catalogue from Table 5 of Bird and Kagan (2004). Events on a boundary may be shared between different zones, which accounts for the fractional numbers. For annual average numbers divide by 25.9

Province	Events 1977–2002.9	M_t (10^{17} Nm)	β	m_C (corner magnitude)	$M_{0\text{total}}$ (10^{20} Nm/ year)	C	Energy (10^{15} J/yr)
Cont. rift boundary	285.9	1.13	0.65	7.64	0.5	0.5	2.5
Cont. transform faults	198.5	3.50	0.65	8.01	1.2	0.7	6.0
Cont. convergent boundary	259.4	3.50	0.62	8.46	3.3	0.9	16.5
Ocean ridge normal faults	424.3	1.13	0.92	5.86	0.31	0.02	1.5
Ocean ridge other mechanisms	77	1.17	0.82	7.39	0.06	0.05	0.3
Ocean transform fault slow	398	2.00	0.64	8.14	2.1	0.9	10.5
Ocean transform fault medium	406.9	2.00	0.65	6.55	0.30	0.1	1.5
Ocean transform fault fast	376.6	2.00	0.73	6.63	0.29	0.1	1.5
Oceanic convergent boundary	117.7	3.50	0.53	8.04	1.5	0.3	7.5
Subduction zones	2052.8	3.50	0.64	9.58	91.3	0.7	456

common value, 2/3 (Table 14.1). Equation (14.44) is empirical and the physical conditions that determine M_C are not fully understood, although geometrical effects, such as the depth of the seismogenic zone, appear to be controlling. Subduction zones have higher values than other regions, especially mid-ocean ridges.

The total moment of the Kagan distribution is

$$M_{\text{total}} = -\int_0^\infty M_0 \mathrm{d}N$$
$$= \int_0^\infty \Big\{ \alpha\beta M_0^{-\beta} \exp(-M_0/M_C) + \alpha/M_C M_0^{1-\beta} \exp(-M_0/M_C) \Big\} \mathrm{d}M_0. \quad (14.45)$$

Noting that the gamma function is given by $\Gamma(z) = \int_0^\infty t^{z-1}\mathrm{e}^{-t}\mathrm{d}t$, and that $\Gamma(z+1) = z\Gamma(z)$, Eq. (14.45) integrates to

$$M_{\text{total}} = \alpha\,\Gamma(1-\beta) M_C^{1-\beta}. \quad (14.46)$$

For $\beta = 2/3$, $\Gamma(1-\beta) = 2.68$, $M_C = 1.2 \times 10^{22}$ Nm with $\alpha \approx 10^{14}$, the average annual global moment release is $M_{0\text{total}} \approx 0.6 \times 10^{21}$ Nm, substantially less than the moments of the Chile 1960 or Sumatra 2004 earthquakes, but these are rare events.

A particular interest in moments is the comparison of the total seismic moment along a plate boundary with that expected from relative plate motion (Brune, 1968). Let M_i be the moment of the ith earthquake and assume that a boundary of length L is slipping to a depth D. Then the total seismic moment for all earthquakes on this area is μLDu where u is the cumulative slip from earthquakes. Seismic coupling, C, is defined as the seismic slip rate u divided by the tectonic rate u_T.

$$u = \frac{\sum M_i}{\mu LD} = \frac{M_{\text{total}}}{\mu LD} = Cu_T, \quad (14.47)$$

where D is the effective depth of the seismogenic zone. u_T is determined from plate reconstructions or from geodetic measurements of current plate motions to ascertain whether all the motion is being taken up by earthquakes or some is taken up by aseismic creep, that is, $C < 1$. It is important that the catalogue extends over sufficient time to be representative of the earthquake activity. While L is easy to measure, D is uncertain. Because earthquakes with the largest magnitudes have the greatest contribution to the slip it is important to include their effect. However, they are rare. This problem can be overcome by fitting the distribution of moment from the catalogue to the Kagan distribution (Eq. (14.44)) to find the corner moment and earthquake rate, and then using Eq. (14.46) for total seismic moment. From a global survey, Kagan (2002a, b) found that on average the worldwide coupling coefficient is $C \approx 0.5$.

The Harvard moment tensor catalogue (Dziewonski *et al.*, 1981) contains seismologically determined moment tensors for earthquakes worldwide since 1977, and is thought to be complete above magnitude 5.6 (threshold moment $M_t = 3 \times 10^{18}$ Nm). Let N_t be the number of events per unit time above the threshold. Then from Eq. (14.44),

$$\alpha = N_t M_t^{\beta} e^{\frac{M_t}{M_C}}. \tag{14.48}$$

Substituting Eq. (14.48) in Eq. (14.46), the total moment becomes

$$M_{0\text{total}} = N_t M_t^{\beta} e^{\frac{M_t}{M_C}} \Gamma(1-\beta) M_C^{1-\beta}. \tag{14.49}$$

and Eq (14.47) gives

$$C = \frac{1}{\mu L D u_T} N_t M_t^{\beta} e^{\frac{M_t}{M_C}} \Gamma(1-\beta) M_C^{1-\beta}. \tag{14.50}$$

Bird and Kagan (2004) have used this model to examine parameters in Eq. (14.47) and coupling as functions of tectonic province. They divided the Harvard moment tensor catalogue into different regions, such as subduction zones, ocean spreading centres and continental convergence zones, as in Table 14.1, and determined N_t, β, M_C, and $M_{0\text{total}}$ from the moment distributions with an estimation of C by Eq. (14.50). The continental collision zones are particularly well observed, with the Alpine–Himalayan chain as the major component. They estimated a total boundary length of 12 516 km and average slip of 5.26 cm y^{-1}, taking into account dip angles and slip on variously oriented faults. If the depth of the seismogenic zone is taken as 20 km, the tectonic moment rate is $\mu L D u_T = 3.65 \times 10^{20}$ Nm y^{-1}, assuming $\mu = 27.7$ GPa. The seismic moment rate was obtained by fitting Eq. (14.49) to the moment tensor data for the years 1977–2002.9 giving $M_C = 5.82 \times 10^{21}$ Nm, $\beta = 0.62$, $m_C = 8.6$, $M_t = 3.5 \times 10^{17}$ Nm, $N_t = 10.01/\text{y}$. With these parameters the annual total seismic moment, for the Alpine–Himalayan chain, is $M_{0\text{total}} = 3.29 \times 10^{20}$ Nm y^{-1}. In this case the ratio of the seismic and tectonic rates (3.29/3.65) gives a coupling coefficient of 0.9. Of the other provinces examined, ocean spreading ridges had the lowest corner magnitudes and coupling ($m_C = 5.86$, $C = 0.02$), which might be expected because they are hot, favouring aseismic creep. By contrast, subduction zones are well coupled ($m_C = 9.6$, $C = 0.7$).

We use these ideas to calculate the global energy release in the various tectonic zones (Table 14.1). Using Eqs. (14.36) and (14.38),

$$E = 5.0 \times 10^{-5} M_0 \, \text{J}. \tag{14.51}$$

Then from Eq. (14.49) we obtain

$$E_{\text{total}} = 5 \times 10^{-5} N_t M_t^{\beta} e^{\frac{M_t}{M_C}} \Gamma(1-\beta) M_C^{1-\beta}. \tag{14.52}$$

Summing the energies in Table 14.1, the average total energy release is 5.04×10^{17} J y^{-1}. We assign an average ‘efficiency’ of 6% to the conversion of strain energy to seismic waves in Chapter 15, making the annual dissipation of strain energy about 8.4×10^{18} J y^{-1} or 2.7×10^{11} W. This is 3.5% of the total convective power of the mantle (Section 22.5). The fraction of mantle volume that is subject to earthquakes is much smaller than 3%, so there is a strong concentration of dissipation in the subduction zones. These numbers confirm that earthquakes are hiccups at sticking points in mantle convection. The average annual energy release corresponds to that of a single magnitude 8.6 event and is exceeded by each of the largest shocks (Chile, 1960; Alaska, 1964; Sumatra, 2004).

14.8 Tsunamis

There are occasional reports of earthquakes having been felt at sea, but the secondary effect of wave generation by submarine earthquakes is much more significant. The waves sweep across open ocean at high speeds and cause severe damage and loss of life on coastal areas many thousands of kilometres from the earthquakes that generate them. These waves were once referred to as 'tidal waves' but are not related to tides and the Japanese word 'tsunami' (literal translation: harbour wave) is now used universally. Much of the perimeter of the Pacific Ocean is seismically active and it has received most attention. This was deemed to have paid off after the successful prediction of the damaging tsunami from the May 1960 Chilean earthquake. However, all oceans are at risk, as the devastating Indian Ocean tsunami triggered by the 2004 Sumatra–Andaman earthquake, reminded us. The principles of tsunami propagation, as reviewed by Stevenson (2005), are reasonably well understood, being an application of classical hydrodynamics, although with complications in detail arising from complexity of sea floor topography.

Earthquakes are not the only cause of tsunamis. The 1883 eruptions of Krakatoa volcano, in the Sunda Strait between Java and Sumatra, produced several tsunamis, the final one being locally very big and destroying hundreds of coastal villages in western Java and southern Sumatra. Slumping of marine sediments also generates tsunamis and, in at least some cases, the slumping is triggered by earthquakes, leading to confusion over the mechanism of tsunami generation. A recent example that drew attention to the problem is the tsunami that struck the Aitape area of New Guinea in July 1988, following a magnitude 7.0 earthquake north of the island. The wave height, peaking close to 15 m, appeared too great to be explained by the earthquake itself and is attributed to a sediment slump triggered by the earthquake. It appears possible that slumping contributes to most major tsunamis, even if they are clearly identified with earthquakes. Geological evidence of very big prehistoric tsunamis links them to sediment slumping, but with no indication whether they were earthquake-triggered. There is an area of massive slumps between Norway and Iceland, each involving more than 1000 km^3 of sediment that slipped hundreds of kilometres into the deep Atlantic from the continental slope off Norway. Best documented is an isotopically dated event about 7000 years ago, coincident with widespread tsunami deposits along the Norwegian and Scottish coasts. The Hawaiian islands also deposit thick sediments on steep offshore slopes and are therefore a likely source of slump-generated tsunamis. A major slump south of the island of Lanai about 100 000 years ago is postulated to have caused a tsunami still evident as marine deposits on the east coast of Australia.

Earthquakes such as Chile, 1960, Alaska, 1964 and Sumatra, 2004, that are major tsunami generators, cause movements of the sea floor that are hundreds of kilometres in extent and so generate waves hundreds of kilometres in wavelength. They are shallow-water waves, meaning that water depth is much less than the wavelength The general equation for the speed of a wave of wavelength $\lambda = 2\pi/k$ in water of arbitrary depth h is

$$v = \left[\frac{g}{k}\tanh(kh)\right]^{1/2}, \tag{14.53}$$

where g is gravity (see, for example, Proudman, 1953). The definition of a shallow-water wave is $(kh) \ll 1$, in which case Eq. (14.53) reduces to

$$v = \sqrt{gh}. \tag{14.54}$$

In water 5 km deep the speed is 220 m s^{-1} (800 km/hour). Although this is very fast by the standards of ocean waves, it is twenty times slower than Rayleigh waves and forty times slower than P-waves, making possible the tsunami warning system that operates in the Pacific. The waves are also slower than the usual speed of fault propagation (by a factor $\sim$10), so that, to a good approximation, tsunami generation by an earthquake is synchronous over the fault area. By Eq. (14.54) the wave speed is depth dependent but not frequency dependent.

An understanding of tsunamis was one of the targets of the International Geophysical Year (IGY), a period of focussed international collaboration in 1957–8. The rarity of major tsunamis had restricted progress and a new series of sensitive recorders was developed, with hydraulic tuning that made them sensitive to waves with periods between 3 minutes and 2 hours while almost ignoring both tides and normal wind-driven waves. They were deployed around the Pacific with the expectation that numerous small tsunamis would be seen, yielding much more data in a short time than would otherwise be possible. But the IGY period was virtually tsunami-free. Small tsunamis are not as frequent as had been supposed. To see a reason for this we compare the energy of a tsunami with that of the earthquake that generates it.

First, we note that earthquakes on faults that are not very large (in both dimensions), compared with the depth of water below which they occur, cannot be significant tsunami generators because the integral of sea floor displacements is zero and adjacent movements of opposite signs cancel within the water column. If we consider an earthquake with mean displacement b across a fault of length L and width W in a medium of rigidity μ then its moment is $M_0 = \mu bLW$ (Eq. 14.6). We can relate it to earthquake energy, E_S, by combining Eqs. (14.36) and (14.38), assuming $M_W = M_S$,

$$E_S = M_0 \times 10^{-4.3}. \qquad (14.55)$$

This effectively assumes that seismic strain is independent of magnitude, which is found to be a satisfactory approximation in Section 15.6, at least for the large shocks that are relevant to tsunamis. Thus, seismic energy is proportional to the product of dimensions, bLW. But, with the assumption of constant strain, b is proportional to the smaller fault dimension, W, so that we have

$$E_S \propto LW^2. \qquad (14.56)$$

The sea floor displacement is some fraction, f, of the fault displacement, b, depending on the inclination of the fault, and, since the fault movement is rapid compared with ocean wave propagation, the movement of the water column is almost synchronous everywhere and follows the sea floor motion. On average it is balanced, that is the total of movements up and down is zero, and the average gravitational energy imparted to the water column is proportional to $(fb)^2$. This becomes the tsunami energy, for which the total is, with substitution of the proportionality of b to W,

$$E_T \propto \rho g (fb)^2 LW \propto \rho g f^2 LW^3. \qquad (14.57)$$

From Eqs. (14.56) and (14.57) we see that, for faults that are equidimensional ($L \approx W$), E_T is proportional to W^4 but E_S is proportional to W^3. This appears to be appropriate for small to moderate shocks and demonstrates that tsunami energy is more dependent on size than is the earthquake energy itself. We can see why small tsunamis are rarer than might have been expected from the distribution of earthquake magnitudes. For the largest shocks (magnitudes 8 to 9+), fault dimension ratios (L/W) are observed to increase with magnitude; in the approximation that W is constant for this size range and only L increases with magnitude, the two energies are proportional to one another. The rare very large shocks are disproportionately effective tsunami generators.

An IGY tsunami recorder was maintained on Norfolk Island, 1200 km east of mainland Australia, for several years after the IGY period and was operating at the time of the 1960 Chile earthquake. For much of the resulting record the instrument was driven off scale (drawing square waves), but satisfactorily recorded the first hour (1.5 cycles) of the arriving wave (Fig. 14.17(a)). The interest in this record arises from the fact that it was obtained at a site that is about as near as one can get to an open ocean situation: an instrument mounted on an exposed jetty on an island roughly 7 km square (small compared with the tsunami wavelength), standing in deep water and remote from other land. But the ocean floor between Chile and Norfolk Island is far from featureless and the depth dependent wave speed means that the arriving wave had been refracted, diffracted and reflected in complicated ways, just as seismic waves are affected by irregularities in the mantle. The lowest frequencies (longest wavelengths) are least affected by the irregularity and so travel faster than the higher frequencies, as in the 'random media' problem in seismology, discussed in Chapter 16. As Fig. 14.17(a) shows, and Fig. 14.17(b) confirms with observations on an estuary near Hilo, Hawaii, the first arriving wave had a distinctly longer period than the following waves, which is what is normally observed, but also that it was relatively small. This observation does not invalidate Eq. (14.54), but it draws attention to the fact that the equation assumes a uniform ocean floor and that for the real ocean there is some dispersion. It means that inferences about source characteristics from wave features must be treated with caution. The 'frosted glass effect' is difficult to allow for.

The problem of tsunami amplitude is of particular concern. If a wave approaches a shore line from a perpendicular direction up a gently sloping sea floor, then the speed decreases by Eq. (14.54), but the wave energy is conserved and the amplitude increases. For a wave of amplitude a and wavelength λ, the energy in one

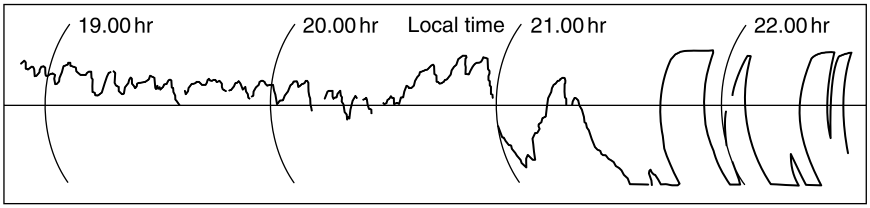


FIGURE 14.17(a) Norfolk Island record of the tsunami from the Chile 1960 earthquake.

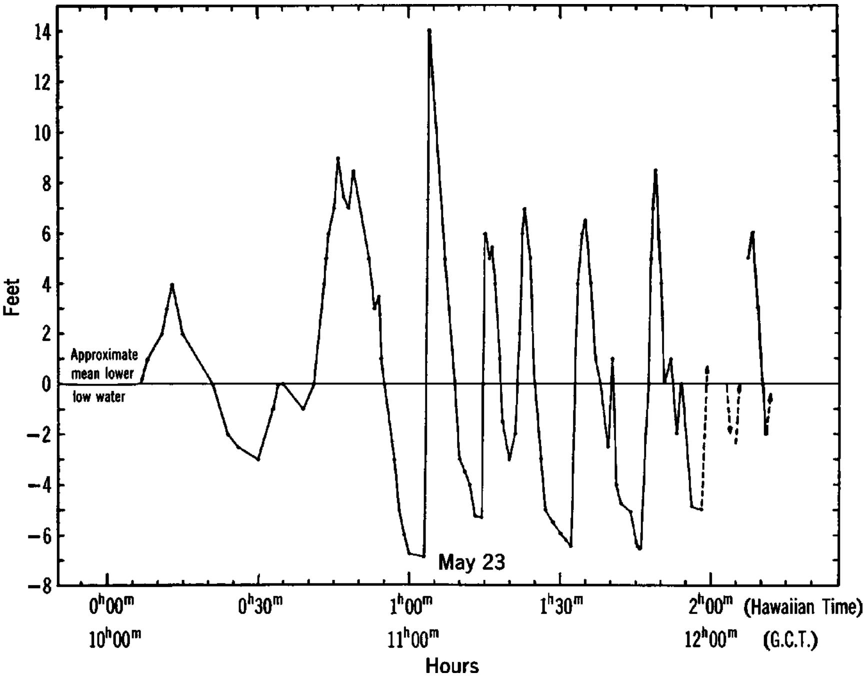


FIGURE 14.17(b) Observations by Eaton *et al.* (1961) of the water level at the Wailuku River bridge, Hilo, Hawaii, during arrival of the tsunami from Chile in May, 1960.

wavelength is proportional to $a^2\lambda$ and since $\lambda \propto V \propto h^{1/2}$, a increases with decreasing water depth h as $h^{-1/4}$. On this basis a 1 m wave in 4 km deep ocean becomes a 4 m wave in 15 m of water, by which point non-linearity is significant and the simple rule breaks down. Other important factors that affect wave amplitude are refraction towards shallow water, which selects certain sections of coastline for enhanced waves according to the sea floor topography, and local resonances (seiches) in bays, estuaries and harbours. A naturally defended coastline is one with sharp submarine contours offshore (perpendicular to the wave direction). The amplitude calculation above assumes a gently sloping sea floor, but if there is a sharp change of depth, relative to the wavelength, then partial reflection occurs. The water motion extends through the full depth of the sea and such depth changes can be modelled as impedance mismatches. A reef, with a steep outer edge distant from the shore, is ideal, blocking motion in much of the water column and reflecting the corresponding fraction of the wave energy.

The floor of the Indian Ocean is subducting beneath Sumatra and the Andaman Islands to the north of it, and the 2004 earthquake was a sharp increment in this process. The subducting (western) plate jerked downwards and the overriding (eastern) plate upwards in the manner illustrated by Fig. 10.14. For waves propagating in the two opposite directions, the first motions, illustrated by a set of five tide records reproduced by Lay *et al.* (2005), were the reverse of what this simple picture would suggest. Similarly, the Norfolk Island record of the Chile tsunami (Fig. 14.17(a)) shows an initial rise in level, although the island is on the subducting side of the fault. In both cases the fault dips were shallow and the simple view of up or down motion on opposite sides is not valid. All of the major tsunamigenic subduction zone earthquakes have focal mechanisms similar to the model of the Alaska earthquake in Fig. 14.7, with shallow dipping fault planes. It is necessary to recognize that the stress release is always of double couple form (Fig. 14.9(b)). As Fig. 14.7 shows, the initial sea floor motion is downwards at the land edge, causing an initial water withdrawal at the nearby shore and initial inundation at remote sites. We need to recognize also that there is probably always some sediment slumping. In the Sumatra case this is consistent with evidence for a very slow component of the tsunami excitation (Lay *et al.*, 2005). The speed of slumping is much slower than the wave speed, so this component behaves quite differently from earthquake excitation and would not have contributed to the first motion of the wave.

14.9 Microseisms

A limit to the useful sensitivity of seismic recording is imposed by background vibration, to which the term microseisms is applied. It is most serious at periods between 5 s and 12 s. The word does not imply that microseisms are caused by numerous small earthquakes. There are several causes, some of which are local, such as the shaking of trees by wind (forests are seismically noisy). Of greater interest is the generation of microseisms by storm waves at sea. Storm-generated microseisms are observed at great distances. They are principally Rayleigh waves (vertically polarized surface waves – see Section 15.3).

Ocean waves are caused by friction between the wind and sea surface. The amplitudes and wavelengths grow with wind velocity and its duration up to a limit. This limit occurs when the phase velocity of the waves is approximately equal to the wind speed, a condition that gives rise to what is called a fully developed sea. For deep water waves ($kh \gg 1$, $\tanh(kh) \to 1$), Eq. (14.53) becomes $V = (g/k)^{1/2}$, so that the period is

$$T = 2\pi V/g. \tag{14.58}$$

For 30 to 40 knot winds (15.4 to 20.6 m s^{-1}), typical of storms, Eq. (14.58) gives periods of 10 to 13 seconds and wavelengths of 150 to 270 m.

Consider a water wave propagating across deep ocean. The particle motion is circular, in the sense of a wheel rotating backwards relative to the forward motion of the wave, giving the surface a trochoidal form, with crests sharper than troughs. But the vertical motion of the

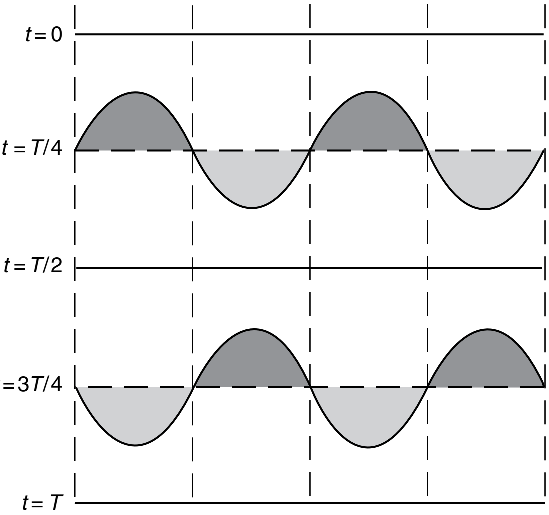


FIGURE 14.18 A series of sea surface profiles at intervals of a quarter of the period, T, of a standing wave. Relative to the flat surface at $t = 0$, $T/2$ and T, water is removed from troughs (lightly shaded) and added to the crests (heavily shaded) twice in each wave period.

water is sinusoidal and so there is a sinusoidal oscillation in pressure below the surface. The amplitude of the pressure oscillation decreases exponentially with depth on a scale determined by the wavelength, because at increasing depth there is an increasingly effective cancellation of contributions by points on the wave that are out of phase. Thus, there is no pressure pulsation on the ocean floor and no microseism generation by propagating waves with wavelengths that are short compared with the depth.

When two similar waves travelling in opposite directions interfere, a standing wave is set up and in this situation a pressure oscillation may be transmitted to indefinitely great depths. A simple version of the theory is given by Longuet-Higgins and Ursell (1948). Consider the sequence of surface profiles in Fig. 14.18. The total gravitational potential energy of the water is greater at $t = T/4$ and $3T/4$ than at the instants when the sea surface is flat. There is therefore a synchronous pressure oscillation of period $T/2$ over the whole of the area over which the standing wave is coherent. If the dimensions of the area are large compared with the depth of water, then the pressure oscillation is transmitted to the sea floor. The effect is a second order one. The frequency of the pressure cycle is twice the frequency of the surface wave and its amplitude is proportional to the square of the wave amplitude. To see why this is so, note that the mass of water that is moved (represented by the cross-section of the shaded areas in Fig. 14.18) and the average height of the movement are each proportional to wave amplitude and that the product of these gives the integrated mass-times-vertical acceleration of the water column and therefore the variation in bottom pressure.

Microseisms may be caused also by storm waves breaking on shorelines. Ocean waves are refracted so that they are more or less parallel to a coastline when they break. This gives them greater coherence and more effective microseism generation than the same energy distributed irregularly. In this case there is no frequency doubling. Storm microseisms originating at shorelines may be distinguished from deep-ocean microseisms by their dominant periods, 10 to 12 s instead of 5 to 6 s. Although the coastal mechanism does not require the special condition of crossing wave-trains and is therefore a more common occurrence, it is the deep-ocean effect that produces the largest microseisms. However, even vigorous storms are not necessarily strong microseism generators. Large waves produced by steady monsoonal winds are not effective. Microseism generation requires cyclonic disturbances that move about and so produce crossing wave-trains.

15

Earthquake dynamics

15.1 Preamble

After more than a century of scientific observation, earthquakes have given a detailed picture of the structure of the Earth, but only a rudimentary idea of the physical processes in the fault zones where they occur. We know that earthquakes are local hiccups in convectively driven tectonic motion and the pattern of their occurrences is a central component of the theory of plate tectonics. Elasticity theory explains the geometry of stress release during earthquakes but not the mechanism that triggers them. Although this depends on the physical properties of materials in the fault zones, we are still far from a level of understanding that might lead to reliable prediction. It is possible that the detailed information that would be required is beyond reach. Nevertheless prediction has been the ultimate target of much seismological research. This chapter is concerned with the information and ideas that the work has produced, with a brief discussion of the possibility of prediction itself in the final section.

A basic problem is that earthquakes are very diverse phenomena. The range of sizes is extremely wide, probably bounded at the upper end of the scale by magnitudes not much greater than that of the Chile 1960 event (Section 14.7) and with no observed lower bound. The range of energies exceeds 10^{17}:1. The range of speeds of fault movement during earthquakes is also very wide, bounded at the upper end by the accelerations that can be produced by stresses of order 10 MPa if fault friction suddenly vanishes. Most well-studied earthquake ruptures have speeds of 2 to 3 km/s, near to the S-wave speed. There are also ‘slowquakes’ that may occur over hours to months and can be regarded as creep events. There is probably no lower bound short of the speed of plate motion, centimetres per year. Seismological observations have emphasized the more rapid events because they radiate elastic waves. The energy release is converted to seismic waves with dissipation by local friction and breakage of bonds. The diversity and wide range of phenomena require an explanation in terms of material properties and fault behaviour. A phenomenological theory, known as the rate–state theory, is the most successful of the current alternatives.

For most earthquakes more energy is released in the fault zones themselves than is radiated and for ‘slowquakes’ virtually no energy at all is radiated. This implies strong heating of fault zones, so that the absence of geothermal anomalies along fault zones requires an explanation. Removal of heat by flow of ground water can be invoked, but invites doubt about the validity of heat flow measurements.

In pondering the problem of earthquake triggering, basic questions are: What distinguishes large from small events? Are they self-similar with only a scale difference and if so what stops them from growing further? Or can we suppose that large events are just superpositions of numerous small ones – ‘instantaneous aftershocks’? The really big events, Sumatra, 2004 being an impressive example, start at a particular point and spread, often in one direction. From a study of this event by Ishii *et al.* (2005) the duration of fault movement was about ten minutes, which, for the 1600 km fault, implies an average break speed of 2.8 km s^{-1}. A break in one section transfers stress, which breaks the next, and so on. In this situation fault instability may not be a useful concept. An individual section of fault may be nowhere near to breaking on its own but respond to triggering by a neighbour or neighbours. The concept of vulnerability appears to be more relevant. A large earthquake requires a large fault area that is vulnerable to triggering by any small patch within it. This puts the prediction question into context. Detailed prediction (time, place, magnitude) would require, as a starting point, identification and monitoring of unstable patches that may be very small individually, but capable of triggering neighbours, and this appears to be an impossible task. The best that can be expected is recognition of vulnerability, that is, evidence that any earthquake that occurs is likely to grow into a big one. The initiation process itself is almost certainly a small-scale one.

On a longer time scale, aftershocks give an indication of the rate of adjustment of stresses in seismogenic zones. Large shocks are followed by smaller ones that outline the fault zone; typically the largest aftershock is one unit on the magnitude scale smaller than the main shock. But foreshocks are also known and a crucial question is: Can they be recognized as such before the main shock has occurred? A particularly striking example was the occurrence of two $M \sim 8$ events shortly before the $M = 9.5$ Chile 1960 earthquake, which appears to have been preceded also by a very large, deep ‘slowquake’. Repeated patterns of seismicity have often been sought, using a frequency of occurrence vs magnitude relationship discussed in Section 14.7, but no clear conclusion has emerged.

15.2 Stress fields of earthquakes

When an earthquake occurs the crack that forms releases stress in its immediate vicinity by relative motion between the two sides of the fault plane, but increases stress in nearby regions. The slip process starts from a nucleation patch and grows at the rupture velocity, generally observed to be slightly less than the S-wave velocity, that is, about 3 km/s. For small events the slipping zone can be modelled as a growing crack that spreads until it is stopped by a barrier or reaches a region of low stress. Until the stopping point is reached, the whole fault is assumed to be slipping, but that may be only a conclusion based on inadequate observation of very small events. Large earthquakes can be regarded as multiple events of this kind. For very long faults such as the San Andreas, the slipping zone grows to a given size, e.g., a roughly rectangular shaped patch extending over the depth of the seismogenic zone, and travels as a slip-patch disturbance from one end of the fault to the other, or perhaps in both directions, from the initiation. The slipping zone may be made up of a heterogeneous distribution of sub-faults or asperities that trigger each other as the stresses of their seismic waves travel to neighbouring regions.

In idealized models the final slip distribution is one that reduces the stress everywhere on a fault to a value near the dynamic friction limit, though some overshoot may occur before friction brings the fault to a stop. Such a slip distribution can be calculated analytically for elliptical cracks, a circular one being mathematically simplest. The results have several features in common. The slip distribution is elliptical. The stress drop is uniform along the fault but stress has singularities that appear at the boundary, and diminishes as $1/\sqrt{\text{distance}}$ from it. The distribution of the dislocation $b(x')$, that is required to give constant stress drop is found by solving the integral equation

$$\Delta\sigma(x) = \text{constant} = \iint_{S'} b(x')G_\sigma(x, x')\mathrm{d}S' \quad (15.1)$$

where $G_\sigma(x,x')$ is the Green's function for stress at x due to a displacement at x' and may be obtained by differentiating Eq. (14.8) and using Eqs. (11.4) and (11.5) to convert displacements to stress components.

The infinitely long fault can be treated as a special case of an elliptical fault. If displacement occurs in the direction of the long axis, in the sense of a screw dislocation (Fig. 14.4), this is referred to as an antiplane crack, for which stresses, $\sigma(x)$, and the slip distribution, $b(x)$, are (Knopoff, 1958)

$$\begin{aligned} \sigma(x) &= \text{initial stress} + \Delta\sigma\left(\left(1-\frac{c^2}{x^2}\right)^{-1/2}-1\right), \quad |x|>c, \\ \sigma(x) &= \sigma_f, \quad |x| \le c, \end{aligned} \qquad (15.2)$$

$$b(x) = \frac{\Delta\sigma}{\mu}(c^2-x^2)^{\frac{1}{2}}. \qquad (15.3)$$

Figure 15.1(a) plots the stress and displacement for this case. The singularity at each edge of the crack arises because the gradient of displacement becomes infinite. The solution is made more realistic by assuming that the material in the vicinity of the crack edge fails plastically, and that the singularity is rounded off in what is referred to as a slip-weakening zone (Fig. 15.1(b)). After faulting, stresses are decreased along the fault but are increased beyond the edges for distances comparable to the fault dimensions. A feature of this model that is of interest is that the condition for constant stress drop is elliptical displacement (Eq. (15.3)) with the maximum value equal to the strain times the fault dimension. This general behaviour is common to all models.

While observed displacements are approximately elliptical they are much rougher than the theoretical ones of Fig. 15.1. The fault plane may contain over- and undershoots of stress release, resulting in a patchwork of stress concentrations of all shapes and sizes, both within and external to the fault. Even the concept of an individual fault plane is a simplification. A more realistic model describes the faulted region as a zone of sub-faults with a power-law distribution of sizes. Each sub-fault is surrounded by its own stress concentration and aftershocks are delayed response to this stress distribution. Laboratory observations, which indicate that the time to rupture is exponentially dependent on stress (Section 15.3), can explain the long sequences of aftershocks that may continue for years. The sub-fault model explains why aftershocks are concentrated in the fault zone of an earthquake, where reduced stress might be expected. Sub-faults with their associated stress concentrations leave behind very rough stress fields. An explanation of the aftershock delays is explored in Sections 15.3 and 15.5.

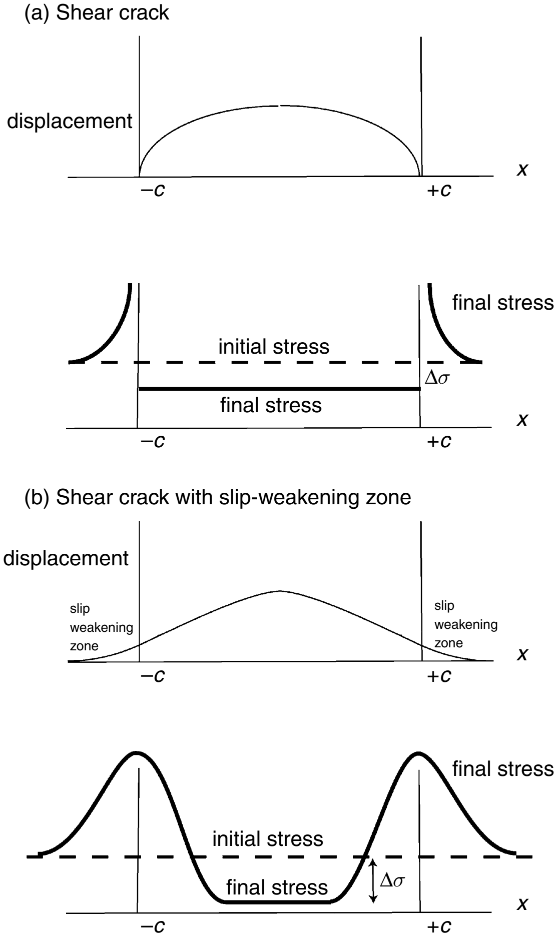


FIGURE 15.1(a) A shear crack, $-c \le x \le c$, with relative displacement of elliptical form, as shown, causes a uniform stress drop on the crack, but singular stress outside, infinite on the boundary. (b) A slip-weakening zone is added to make the model in (a) more realistic, with displacement graded smoothly to zero and all stresses remaining finite.

The distribution of aftershocks in the region of an earthquake has usually been fitted to the pattern of the change in Coulombic stress resulting from the main shock (Stein, 1999). This is given by Eq. (11.36):

$$\tau \geq \Delta\tau_C \geq \mu_S \Delta\sigma_n + S_0, \tag{15.4}$$

where μ_S is the limiting static friction at which sliding begins, $\Delta\sigma_n$ is the normal stress and S_0 is the cohesion. However, this (variation as $1/r^3$) has been observed not to fit the observed $1/r$ variation in frequency of occurrence with distance r from a main shock (Felzer and Brodsky, 2006). It appears possible that aftershocks reflect the distribution of sub-faults, referred to above, that are caused by the dynamic stress of the main shock.

Geodetic measurements following large earthquakes have shown that regional strain continues in the same directions as in the main events with characteristic time constants ranging from months to years. Some post-earthquake strain is expected from the aftershocks, but that is not a complete explanation. The cumulative aftershock moment is generally one to two orders of magnitude less than the geodetically observed moment released after the event. Two models are under investigation. One involves further aseismic slip of the rupture plane itself. The other involves slip in the lower ductile crust resulting from the increased stress external to the fault plane. Recently inversions of post-seismic slip from two large Californian earthquakes (Landers, 1992 and Hector Mine, 1998) have shown that aseismic slip both on the fault and below it is required to satisfy the data. The after-slip may continue for years. Thatcher (1983) estimated a time constant of 30 years for the 1906 San Francisco earthquake. The Maxwell relaxation time (for the model in Fig. 10.4(a)) is given by

$$\tau_M = \eta/\mu. \tag{15.5}$$

If the lower crust is relaxing over 30 years then, for $\mu = 5 \times 10^{10}$ Pa, the viscosity of the lower crust is $\eta = 5 \times 10^{19}$ Pa s. A zone of steady deformation of order 200 km wide in Southern California is seen to be overlaid by patches of more concentrated strain change. Jackson *et al.* (1997) concluded that they can be attributed to earthquake after-slip of historic earthquakes. This complicates any attempt to use strain observations to infer strain build-up for future earthquakes.

Static stresses fall off with distance too rapidly (as $1/r^3$) to cause triggering of remote earthquakes. Dynamic stresses of seismic waves diminish less rapidly, particularly for surface waves which are spreading over a surface and not a volume so that wave energy falls off as $1/r$ and amplitude and stress as $1/\sqrt{r}$. This is why, at teleseismic distances, surface waves dwarf the body waves, for which amplitude decreases as $1/r$. Earthquakes may be triggered by surface waves from distant events, but this is observed to occur only under special conditions, found in hydrothermal or volcanic zones that are characterized by high pore pressure. An example of triggering that appears to have been a prolonged response to dynamic stressing is seen in a several-year period of increased seismicity in the magmatic zone of Long Valley, California, immediately following the Landers 1992 earthquake. Long Valley is several hundred kilometres from Landers and static stress changes from the earthquake would have been negligible.

15.3 Fault friction and earthquake nucleation: the quasi-static regime

The classical description of faulting is that faults break when the driving shear stress is greater than the limiting static friction, and once slip starts the friction drops to a lower value, the so-called dynamic friction. When the reduction in friction with slip is greater than the reduction in the driving stress from the surrounding elastic medium, the fault is unstable and accelerates, causing an earthquake. This stick–slip model of earthquakes was prompted by laboratory observations on friction between rock surfaces.There are several possible reasons for the difference between static and dynamic friction. The presence of fluids introduces several possibilities and is probably important, but there are two mechanisms that do not depend explicitly on fluids. They both represent friction in terms of the interactions between asperities on adjacent surfaces subjected to a normal stress. In the static situation with prolonged contact, asperities become partly welded, and deform, increasing the contact area and consequent friction. With prolonged shear they may move sufficiently to break contact so that new contacts are established with smaller welded areas and this reduces the friction, eventually precipitating an earthquake. This is the essence of the rate–state theory (Dieterich, 1979a, b, 1994). Dynamic friction, which takes over when fault motion is established, involves different mechanisms. Asperities brush past one another rapidly without establishing welded contacts and, in doing so, generate high-frequency elastic waves, making the asperity interactions less effective. With rapid motion fluid lubrication probably also reduces the friction.

We need a dynamic model of earthquakes into which the stick–slip (rate–state) mechanism can be introduced. A convenient analogue is the spring–block slider system (Fig. 15.2). Progressive extension of the spring simulates the tectonic motion that drives the fault movement. The block in the figure moves intermittently, representing earthquakes. The size of the block corresponds to the size of the earthquakes with a 10 km block representing a magnitude 6 event. Let y be the position of the block, and y_{p} the position of the driving plate. The force acting on the block, $k(y_{\mathrm{p}} - y)$, where k is the spring constant, is opposed by the viscous force $\eta\dot{y}$ and the friction F. The equation of motion is

$$m\ddot{y} = k(y_{\mathrm{p}} - y) - \eta\dot{y} - F. \tag{15.6}$$

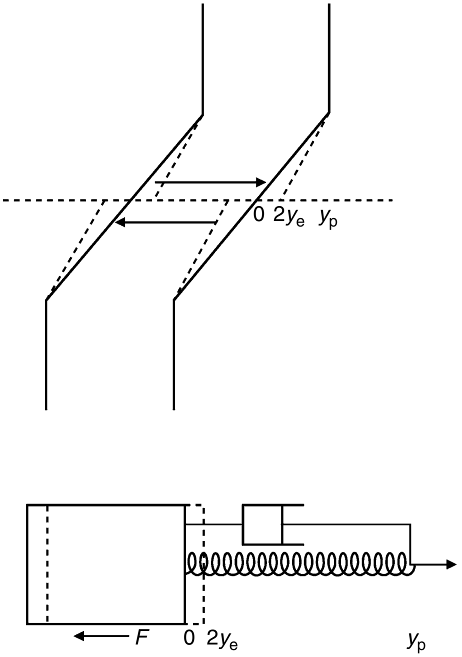


FIGURE 15.2 An earthquake is modelled as two blocks in frictional contact that slide relative to one another. The lower figure shows the spring–block analogue used to investigate the effects of quasi-static and dynamic friction. After a displacement y_e the force exerted by the spring equals the dynamic friction F.

This is the equation for decaying oscillatory motion. There is a transitional phase between static and dynamic friction in the development of instability during which we ignore viscosity and for which there are no inertial or dynamic effects. This is the quasi-static regime. Now we examine the friction term, F, in terms of rate–state theory. Once slipping begins, the change from static to dynamic friction does not take place instantaneously. It requires the surfaces to move a critical distance, D_{c}, to displace interacting asperities with respect to each other sufficiently to erase any memory of the previous regime. D_{c} is taken as the average size of the asperities. For this transitional, quasi-static regime the friction depends on the time asperities are in contact with one another, and therefore on velocity. Since plastic failure has less time to develop than in the static case, we have a quasi-static friction that decreases with velocity as

$$\mu_{\mathrm{Q}} = \mu_0 - a\ln(\dot{y}/\dot{y}_0), \text{ for } \dot{y} > \dot{y}_0, \tag{15.7}$$

where a is a constant, and μ_0 is the friction when the velocity equals the reference velocity $\dot{y}_0$. For the static case, the friction, μ_{S}, arising from the plastic failure (or chemical changes) of individual contacts increases logarithmically with time.

$$\mu_{\mathrm{S}} = \mu_0 + a\ln(t/D_{\mathrm{c}}\dot{y}_0). \tag{15.8}$$

Rabinowicz (1965) unified the concepts of transient and static friction by suggesting that the static friction of two surfaces that have been held together for time t should equal the friction when they are moving at a constant velocity, $\dot{y} = D_{\mathrm{c}}/t$. Then, for those asperities that are in contact in the quasi-static case, plastic flow occurs for the same interval of time, t. Equation (15.7) becomes $\mu_Q = \mu_0 + a\ln(t/D_c\dot{y}_0) = \mu_S$, that is equivalent to the static case, Eq. (15.8), a conclusion that has been confirmed experimentally (Scholz, 1990).

These results can be summarized by the rate–state friction law,

$$\mu_f = \mu_{0f} + A\ln(\dot{y}/\dot{y}_0) + B\ln(\theta) \tag{15.9}$$

where A and B are constants (of order 0.01 and 0.013). The second term, the A-term, represents effects that increase friction with velocity in a manner similar to viscosity. The third term, the B-term, describes the chemical adhesion between the asperities that increases with contact time so that θ is the effective time of contact. Dieterich (1994) shows that the state variable is given by

$$\mathrm{d}\theta = \left[\frac{\dot{y}_0}{\dot{y}} - \frac{\theta}{D_c}\right]\mathrm{d}y. \tag{15.10}$$

For steady-state creep at an arbitrary creep rate $\dot{y}$, $\mathrm{d}\theta/\mathrm{d}y = 0$ and then Eq. (15.10) gives $\theta = D_c\dot{y}_0/\dot{y}$ so that the friction coefficient is

$$\mu_Q = \mu_{0f} + B\ln(D_c) + (A - B)\ln(\dot{y}/\dot{y}_0), \text{ for } \dot{y} > \dot{y}_0. \tag{15.11}$$

Comparing Eq. (15.7) with Eq. (15.11), we see that $a = A - B$ and $\mu_0 = \mu_{0f} + B\ln(D_c)$. If $B > A$ then an increase in $\dot{y}$ causes a decrease in friction. This is an unstable situation, called velocity weakening, which is a necessary preliminary to an earthquake. If $B < A$ this is a stable condition called velocity strengthening. In earthquake regions the boundary between zones of velocity weakening and velocity strengthening is thought to occur at the base of the seismogenic zone marking a change in material properties, controlled by temperature.

At high velocities, still within the quasi-static regime, the first term in Eq. (15.10) becomes small. If it is neglected we have $\theta = \exp(-y/D_c)$ and Eq. (15.9) becomes

$$\mu_f = \mu_{0f} + A\ln(\dot{y}/\dot{y}_0) - By/D_c. \tag{15.12}$$

Using μ_f (Eq. (15.12)) to substitute for F by $F = \mu\sigma S$ in Eq. (15.6), where S is area and σ is normal stress, the equation of motion for the block becomes

$$m\ddot{y} = k(y_p - y) - \mu_{0f}\,\sigma S - A\sigma S\ln(\dot{y}/\dot{y}_0) + B\sigma Sy/D_c \tag{15.13}$$

Since we are considering the quasi-static regime, we assume the inertial term is zero ($m\ddot{y} = 0$) and Eq. (15.13) reduces to

$$0 = -\frac{k}{A\sigma S}(y_p - y) + \frac{\mu_{0f}}{A} + \ln(\dot{y}/\dot{y}_0) - \frac{By}{AD_c} \tag{15.14}$$

which is a differential equation for y with the form

$$0 = \ln(\dot{y}/\dot{y}_0) - Cy + E, \tag{15.15}$$

where $C = \left(\frac{B}{AD_c} - \frac{k}{A\sigma S}\right)$ and $E = \left(\frac{\mu_{0f}}{A} - \frac{ky_p}{A\sigma S}\right)$ are constants. Equation (15.15) is solved by rewriting it as $\dot{y}/\dot{y}_0 = \exp(Cy - E) = \exp(-E)\exp(Cy)$, then multiplying by $\exp(-Cy)\mathrm{d}t$ to get two integrals, $\exp(-Cy)\mathrm{d}y = \dot{y}_0\exp(-E)\mathrm{d}t$, which integrates to $(-1/C)\exp(-Cy) = \dot{y}_0\exp(-E)t + \text{const}$. The constant is found by setting $y = 0$ at $t = 0$, which gives the solution

$$y = \frac{E}{C} - \frac{1}{C}\ln(\exp E - \dot{y}_0Ct). \tag{15.16}$$

Differentiating Eq. (15.16),

$$\dot{y} = \frac{1}{\exp E - \dot{y}_0Ct}. \tag{15.17}$$

The time to instability for $C > 0$ is obtained by setting the denominator in Eq. (15.17) to zero,

$$t_{\text{failure}} = \frac{1}{\dot{y}_0C}\exp\left(\frac{\mu_{0f}}{A} - \frac{ky_p}{A\sigma S}\right) \propto \exp\left(-\frac{\tau}{A\sigma}\right), \tag{15.18}$$

where τ is the shear stress. The essential conclusion from this equation is that the time to failure depends exponentially on the ratio of shear to normal stresses. Alternatively, if $C \leq 0$ this model implies stability, which means failure at negative time, that is in that case the model is recovering from a hypothetical instability in past time.

We now use the rate–state theory to model the behaviour of a fault in the ductile region below the seismic zone. This is the velocity strengthening condition, with $(B < A)$ in Eq. (15.11) (Marone *et al.*, 1991; Hearn, 2003). Consider the blocks in the aseismic zone which are stressed by earthquake displacements of the overlying blocks in the seismogenic zone above them. They can be modelled by a similar spring–block system to that represented in Fig. 15.2. After the upper elastic blocks have slipped in an earthquake they exert a force $k(y_{\mathrm{P}} - y)$ on the lower blocks that then slip aseismically along the boundary separating them, with friction given by Eq. (15.11), that is

$$\mu_{\mathrm{Q}} = \mu_0'' + (A - B)\ln(\dot{y}/\dot{y}_0), \tag{15.19}$$

where μ_0'' is the friction coefficient at the critical velocity $\dot{y}_0$, marking the boundary between stable and unstable states, $(y_{\mathrm{P}} - y)$ is the position of a lower block relative to the one above, and k is the spring constant for the force exerted by an upper block on the lower one. As the lower block moves, the stress decreases. The equation of motion is given by the quasi-static version of Eq. (15.14), which becomes

$$0 = \frac{k}{(A - B)\sigma S}(y_{\mathrm{P}} - y) - \frac{\mu_0''}{(A - B)} - \ln\dot{y} + \ln\dot{y}_0. \tag{15.20}$$

Let the initial velocity, immediately after the earthquake, be V_{c}. Then Eq. (15.20) integrates to

$$y = \frac{A - B}{k}\ln\left[\left(\frac{kV_{\mathrm{c}}}{A - B}\right)t + 1\right]. \tag{15.21}$$

As these blocks move to relieve their own stresses they apply stress to the region above and this is observed as post-seismic strain. Hearn (2003) concluded that Eq. (15.21) gives a satisfactory fit to geodetic observations of post-seismic displacements following several large earthquakes (Landers, Izmit and Hector Mine). This model appears relevant to the effect of the layer immediately below the seismogenic zone, but less so for deeper material which, being hotter, is probably better represented as a viscous medium, as in Eq. (15.5). Long-term relaxation after large events probably includes asthenospheric flow and accounts for large-scale, long-term clustering of earthquakes (Section 15.7).

Harmonic tremor (low-frequency micro-earthquakes) originating at the base of the seismogenic layers of the Cascades subduction zone (Dragert *et al.*, 2001, Rogers and Dragert, 2003) and the San Andreas fault south of Parkfield, California (Nadeau and Dolenc, 2005) is explicable as behaviour of material just marginally in the velocity weakening state. This means a stick–slip type process moderated by viscosity of fluids in the fault plane.

The critical distance D_{c} for establishing a new asperity regime as applied in Eq. (15.8) is not identifiable in earthquake studies. Values of microns to tens of microns are found in laboratory experiments, but seismic observations have too low a resolution for its detection. The asperities in the laboratory samples depend on grain size and on the preparation of slipping surfaces. Natural failure surfaces are much rougher, so that D_{c} is expected to be larger.

The maximum displacement of an equidimensional crack of radius l, responding to a shear stress τ, is obtained from Eq. (15.3), with $\Delta\sigma = \tau$, $x = 0$, $c = l$, as

$$D_{\mathrm{c}} = l\tau/\mu. \tag{15.22}$$

We can relate this to the seismic moment and hence micro-earthquake magnitude that such a displacement would represent. For a stress drop $\Delta\sigma$ in medium of rigidity μ, the moment corresponding to D_{c} is

$$M_0 = D_{\mathrm{c}} l^2 \mu = l^3 \Delta\sigma, \tag{15.23}$$

so that $l = (M_0/\Delta\sigma)^{1/3}$. This imposes a minimum size on the magnitudes of earthquakes that can be triggered by the rate–state process. The Gutenberg–Richter number–magnitude relationship (Eq. (14.39)), is linear down to at least $M_{\mathrm{W}} = 0$ (Section 14.7), corresponding to $M_0 = 1.3 \times 10^9$ Nm. For $\Delta\sigma = 5$ MPa, $l = 6.3$ m and therefore with ambient shear stress $\tau = 30$ MPa and $D_{\mathrm{c}} = 6.3$ mm. The Gutenberg–Richter relationship may extend down further but at such small magnitudes earthquake catalogues are incomplete. If, instead, D_{c} is as small a 100 μm, then the cut-off magnitude is $M_{\mathrm{W}} = -3.6$. Lack of evidence for accelerating strain before earthquakes (Johnston and Linde, 2002; Johnston *et al.*, 2006) indicates that the sizes of the nucleating patches are very small.

15.4 The dynamic regime

Once nucleation is complete and the blocks either side of a fault start to slide, the system passes from a quasi-static to a dynamic regime, and inertial effects in Eq. (15.13) become important. The rate–state mechanism no longer applies and we consider a simplified version of the spring–block model. We neglect viscosity and consider the spring constant, inertia and friction. When the tension in the stretching spring reaches the limiting static frictional force on the block, $F_S = ky_p$, and it starts to move, the frictional force becomes the dynamic friction $F_D = ky_e$ and the block starts to accelerate at a rate $(F_S - F_D)/m = k(y_p - y_e)/m$. At arbitrary y,

$$m\ddot{y} = k(y - y_e). \tag{15.24}$$

This is the equation of simple harmonic motion about y_e, noting that $\ddot{y}$ is positive when $y < y_e$. When the block reaches y_e the net force on it reverses sign and it completes a half cycle about y_e, to reach an extreme position of $2y_e$. At this point it becomes stationary, static friction takes over and there is no further motion even if there is overshoot of the equilibrium position. The solution to Eq. (15.24) is

$$y = y_e(1 - \cos(\omega t)), \tag{15.25}$$

where $\omega = \sqrt{k/m}$. The difference between static and dynamic stress determines how far and how fast a block travels. The rise time of the pulse is the duration of the motion $\tau = \pi/\omega$. The maximum velocity, $V = \omega y_e$, occurs at $y = y_e$, and the maximum acceleration, $a = \omega^2 y_e$, occurs at $y = 0$.

We put in some numbers for illustration, translating the block model parameters to seismic parameters. Let the limiting static stress be $\sigma_S = ky_p/L^2 = 30\,\text{MPa}$ and the dynamic stress $\sigma_D = k(y_p - y_e)/L^2 = 25\,\text{MPa}$. The final stress is $\sigma_F = k(y_p - 2y_e)/L^2 = 20\,\text{MPa}$, giving a stress drop $\Delta\sigma = (\sigma_S - \sigma_F) = 2(\sigma_S - \sigma_D) = 2ky_e/L^2 = 10\,\text{MPa}$. For shear modulus μ, stiffness is given by $k = \mu L$, where L is the block dimension. From Eq. (15.25) the equilibrium position is $y_e = \Delta\sigma L^2/2k = \Delta\sigma L/2\mu = 1.5$ m; the maximum velocity is $V = \Delta\sigma/\mu\sqrt{\mu/\rho} = \varepsilon V_S$, where ε is maximum strain and V_S is the shear wave velocity. These relations are listed in Table 15.1, which gives

Table 15.1 Spring–block analogue for a 10 km × 10 km earthquake ($M \approx 6$)

Block dimension	L	10 km
Shear modulus	μ	3.3×10^{10} Pa
Density	ρ	3×10^3 kg
Shear wave velocity	$V_s = \sqrt{\mu/\rho}$	3.3 km s^{-1}
Stiffness	$k = \mu L$	3.3×10^{14} Nm^{-1}
Limiting static friction	σ_S	30 Mpa
Dynamic friction	σ_D	25 Mpa
Stress drop	$\Delta\sigma = (\sigma_S - \sigma_F)$	10 Mpa
Maximum strain	$\varepsilon = \Delta\sigma/\mu$	3×10^{-4}
Equilibrium position	$y_e = \varepsilon L/2$	1.5 m
Maximum displacement	$2y_e$	3 m
Rise time	$\tau = \pi/\omega$	9.5 s
Maximum velocity	$V = \omega y_e = \varepsilon V_s/2$	0.5 m s^{-1}
Maximum acceleration	$a = \omega^2 y_e = \varepsilon V_s^2/2L$	0.16 m s^{-2}
Radiation efficiency (Eq. (15.52))	$\eta_R = \dfrac{1}{40}$	2.5%
Seismic efficiency (Eq. (15.51))	$\eta_S = \dfrac{\Delta\sigma}{40(\sigma_S + \sigma_D - \Delta\sigma/2)}$	0.5%

values in the range of those observed for earthquakes of this magnitude. This is of course a simplified model in which the block is moving over an inertially fixed plane. The simple frictional model predicts that earthquakes should be periodic and not irregular as is observed. This is a shortcoming of the assumptions of the simple frictional model and is the essential reason for considering more complicated frictional theories such as rate–state.

15.5 Omori's aftershock law

Omori (1894) found that the rate of occurrence of aftershocks, $R(t)$, as a function of time, t, after a mainshock decays as inverse time (Fig. 15.3),

$$R(t) = A/t. \tag{15.26}$$

This law was generalized by Utsu (1961) to

$$R(t) = A\frac{1}{(t+c)^p}, \tag{15.27}$$

with $p \approx 1$, and c a constant, avoiding the singularity at $t = 0$. A law of this type, where t is the time to the mainshock, has also been used to represent foreshocks for those events for which they are clearly identified.

Such power-law clustering has been described by a number of physical mechanisms that have in common the observation that failure is not an instantaneous process, but develops on a time scale that is non-linearly dependent on stress. They include fault healing, static friction increasing with time (the rate–state theory) and stress corrosion, and it is possible that several operate simultaneously. The principle can be

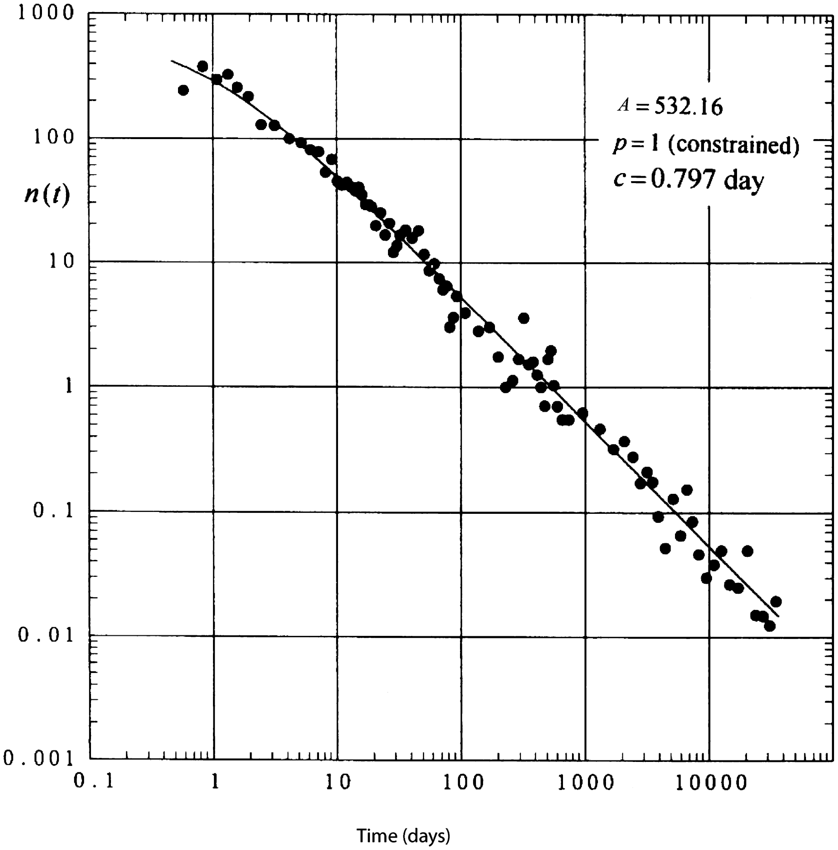


FIGURE 15.3 The decay of aftershock activity after the 1891 Nobi, Japan, magnitude 8 earthquake, fitted to Eq. (15.27). $n(t)$ is the number of shocks per day, identified as aftershocks, which are seen to have persisted for 80 years. Figure from Utsu (2002).

illustrated by examining equations that have been used to describe the stress corrosion mechanism. Material subjected to high stress, especially at the tips of cracks, is vulnerable to corrosion by fluids because the atomic bonds are weakened by the stress. Two alternative expressions have been used to describe the time to failure of an aftershock fault in terms of stress corrosion theory:

$$t_{\text{failure}} = t_1 \tau^{-n} \exp(\Delta H/RT) \tag{15.28}$$

and

$$t_{\text{failure}} = t_2 \exp[(-\beta\tau + \Delta H)/RT], \tag{15.29}$$

where t_1 and t_2 are constants, τ is stress, ΔH is an activation enthalpy and n and β are coefficients that are fitted to observations (Atkinson, 1982; Meredith and Atkinson, 1983). As we saw in Section 15.3, the rate–state model gives a stress dependence with the form of Eq. (15.29), where temperature is assumed constant, and we examine the application of this equation.

Equation (15.29) can be written

$$t = t_0 \exp(-\lambda\tau). \tag{15.30}$$

The number of patches with shear stress between τ and $\tau + \mathrm{d}\tau$, $S(\tau)\mathrm{d}\tau$, is equated to the number with failure times between t and $t + \mathrm{d}t$, $R(t)\mathrm{d}t$

$$R(t)\mathrm{d}t = -S(\tau)\mathrm{d}\tau. \tag{15.31}$$

Substituting from Eq. (15.30) we obtain

$$R(t) = -S(\tau)\frac{\mathrm{d}\tau}{\mathrm{d}t} = \frac{S(\tau)}{\lambda t_0}\exp(\lambda\tau), \tag{15.32}$$

$$R(t) = \frac{S(\tau)}{\lambda t}, \quad \tau < \tau_0, \tag{15.33}$$

where τ_0 is an upper bound, the threshold stress for plastic failure, caused by slip weakening. This bound disallows infinite stress, requiring $S(\tau \to \infty) = 0$ in Eq. (15.33). Equation (15.33) gives the $1/t$ variation of Omori's law, provided that $S(\tau)$ is a weak function of τ compared with the exponential, which makes $R(t)$ a strong function of t. Imposition of the stress threshold can account for the round-off in the Omori law at short times. Shaw (1993) presented a similar analysis using Eq. (15.28).

15.6 Stress drop and radiated energy

One of the fundamental problems of earthquake dynamics is the fraction of the released elastic energy that goes into seismic waves, and how much is dissipated in the fault as heat and the breakage of new surfaces. Consider a fault of area S across which slip b occurs during an earthquake. If the stress across S is initially σ_1 and decreases during the earthquake to σ_2, then the total energy release by the earthquake is

$$E_{\text{total}} = \frac{1}{2}(\sigma_1 + \sigma_2)Sb. \tag{15.34}$$

We distinguish two components of E_{total}, the energy radiated as elastic waves, E_{R}, and the balance that is dissipated in the fault zone. The breakage of rocks in the vicinity of the fault and the production of powdered fault gouge is generally recognized, and conventional frictional heating, even to the point of forming a thin layer of melt, has been inferred from exhumed faults. However, seismic wave observations cannot distinguish heating from rock breakage. We represent the total dissipation in the fault zone as D. Then

$$E_{\text{total}} = E_{\text{R}} + D. \tag{15.35}$$

The essential difficulty in apportioning the energy partitioning is that seismic observations do not give values of σ_1 or σ_2 but, as we show below, give their difference $\Delta\sigma = \sigma_1 - \sigma_2$, which is referred to as the stress drop. E_{R} is found as elastic wave energy from seismograms. There is no direct observation of E_{D}, and E_{total} can be estimated, where observed displacements can be fitted to dislocation models, only with assumptions about σ_1 and σ_2.

There are two distinct quantities, referred to as efficiencies, that are ratios of energies, illustrated in Fig 15.4. The seismic efficiency, η_{S}, is the ratio of radiated to total energy

$$\eta_{\text{S}} = \frac{E_{\text{R}}}{E_{\text{total}}}. \tag{15.36}$$

This can be determined if σ_1 and σ_2 are estimated independently of seismic wave observations. We

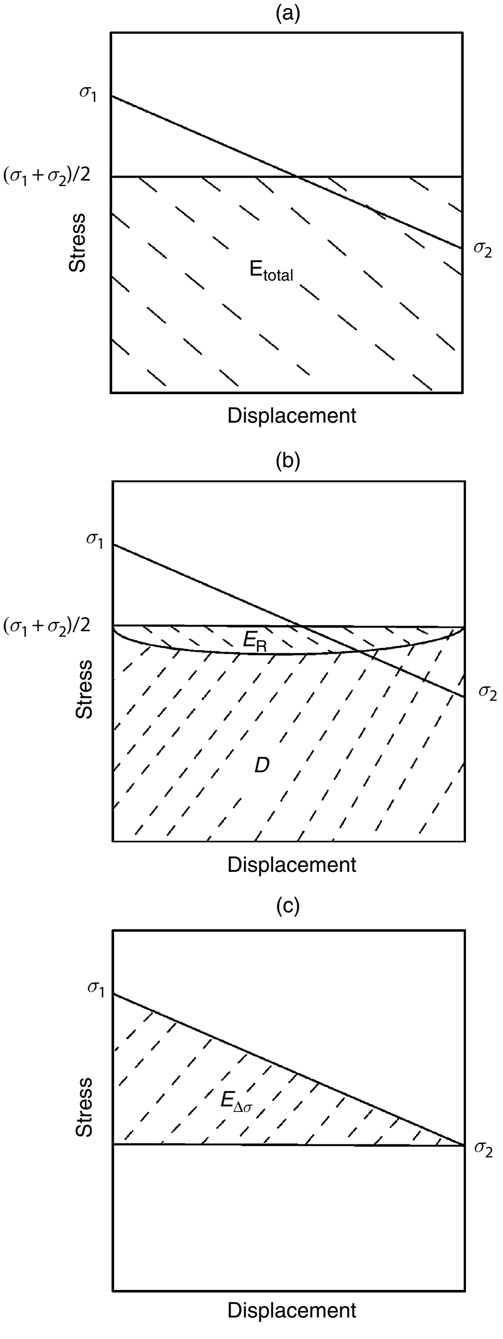


FIGURE 15.4 Energy terms used in defining seismic and radiation efficiencies. Stress during an earthquake drops from σ_1 to σ_2, taken here as linear with displacement, but it can have a more complicated variation. (a) Total energy. (b) Radiated energy, E_R, and energy dissipated as frictional heat and rock breakage, D. (c) Stress drop energy, $E_{\Delta\sigma}$.

consider this further below to compare E_{total} with the tectonic energy budget. We now consider the second efficiency, η_R, termed radiation efficiency, that can be estimated from seismic waves without additional information. We identify with the stress drop a notional energy, $E_{\Delta\sigma}$, that can be estimated from the moment and stress drop if a characteristic dimension of an event is determined from either its spectrum, geodetically observed displacements, or, for large events, the distribution of aftershocks. Then the radiation efficiency is

$$\eta_R = \frac{E_R}{E_{\Delta\sigma}}. \tag{15.37}$$

Values of about 50% are typical, but have a wide range (Kanamori and Brodsky, 2004).

Recalling that seismic moment (Eq. (14.6)) is

$$M_0 = \mu S b. \tag{15.38}$$

the elastic energy, $E_{\Delta\sigma}$, identified with a stress drop $\Delta\sigma$ is

$$E_{\Delta\sigma} = \frac{\Delta\sigma}{2\mu} M_0. \tag{15.39}$$

M_0 is estimated from the zero-frequency asymptote of the spectrum of seismic waves (Section 14.5); $\Delta\sigma$ can be calculated from the moment if we can estimate a characteristic fault dimension, l, since for an equi-dimensional fault

$$\Delta\sigma = M_0/l^3. \tag{15.40}$$

To obtain l we use the corner frequency, ω_0 (Section 14.3), for which there is a direct relationship with the duration of the rupture process, so that we can write

$$\omega_0 = \zeta V_R/l, \tag{15.41}$$

where V_R is the speed of propagation of rupture, and the numerical constant $\zeta \approx 3.5$ is somewhat model-dependent. Observations give

$$V_R \approx 0.9 V_S, \tag{15.42}$$

that is, a rupture spreads at about 90% of the speed of shear waves in the rupturing medium. Equations (15.41) and (15.42) can be used to find l from measurements of moment and corner frequency. There are numerous determinations of stress drop, and for earthquakes of magnitudes exceeding 3 or 4 the universal conclusion is that $\Delta\sigma$ is centred on the range 1 to 10 MPa (10 to 100 bar), as in Fig. 15.5(a). Thus, assuming the stress drop to be approximately a universal constant (at least for earthquakes with $M_W > 3$, $\Delta\sigma/\mu = \varepsilon \approx 10^{-4}$), the moment (or magnitude) is determined by the area of the fault that breaks,

$$M_0 \approx \Delta\sigma S^{3/2}. \tag{15.43}$$

The observation that $M_0 \propto S^{3/2}$ for moderate to large events (Henry and Das, 2001; Kanamori and Brodsky, 2004) has provided the strongest evidence that stress drops have a limited range.

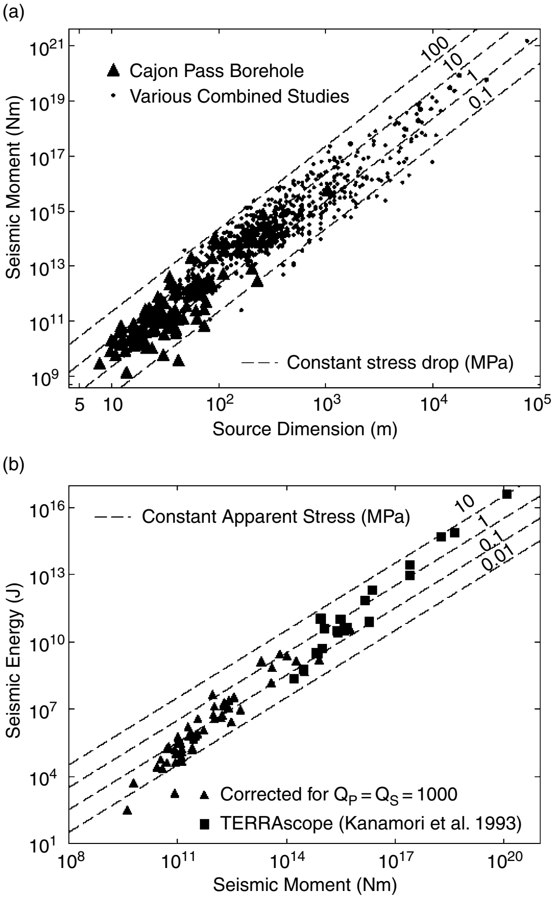


FIGURE 15.5 (a) Seismic moment versus source dimension from various studies compiled by Abercrombie and Rice (2005), including small events detected by a low-noise seismic station at the base of a borehole in Southern California (Cajon Pass). Stress drops calculated from $M_0 = \Delta\sigma l^3$ (Eq. (15.40)), plotted as dashed lines, are centred on the range 1–10 MPa. (b) Radiated energy versus moment estimates by Abercrombie and Rice (2005). Corresponding apparent stresses calculated from $\mu E_R/M_0$ (Eq. (15.44)) are plotted as dashed lines. Disregarding the smallest shocks, the data are centred on the range 0.1–1 MPa. Stress drop is about three times the apparent stress. Thus, by Eq. (15.45), $\eta_R = 2\sigma_a/\Delta\sigma$ and radiation efficiency is 0.6.

Another term, in common usage, that can be estimated from observations is apparent stress, σ_a, defined as

$$\sigma_a = \frac{\mu E_R}{M_0}, \tag{15.44}$$

which is generally found to be lower than stress drop, as can be seen by comparing Figs. 15.5(a) and (b). Combining Eqs. (15.37), (15.39) and (15.42) we see that the radiation efficiency is

$$\eta_R = \frac{2\sigma_a}{\Delta\sigma}. \tag{15.45}$$

Values of η_R range from a few per cent to 100% (Fig. 15.6), with the higher values corresponding to higher rupture velocities. For small earthquakes, constant $\eta_R = 0.66$ appears to be a better approximation than constant stress drop (Ide *et al.*, 2003).

Seismic radiation has near-field terms, which decay rapidly with distance from an exciting earthquake and include static terms. They are important to the strong ground motion near to an earthquake, but for the present discussion we are more interested in the far-field radiated P or S energy from a dislocation (Aki and Richards, 2002), which depends on the acceleration of the moment as

$$E_R^P = \int_0^\infty \frac{\ddot{M}_0^2 dt}{15\pi\rho V_P^5}, \tag{15.46}$$

$$E_R^S = \int_0^\infty \frac{\ddot{M}_0^2 dt}{10\pi\rho V_S^5}. \tag{15.47}$$

Since $V_S^5/V_P^5 \ll 1$, the radiated energy is dominated by the S-waves. A very slow earthquake with $\ddot{M}_0 \approx 0$ generates negligible radiated energy (e.g., Fig. 15.6).

Using the block model from Section 15.4, we can evaluate Eq. (15.47) using the relations in

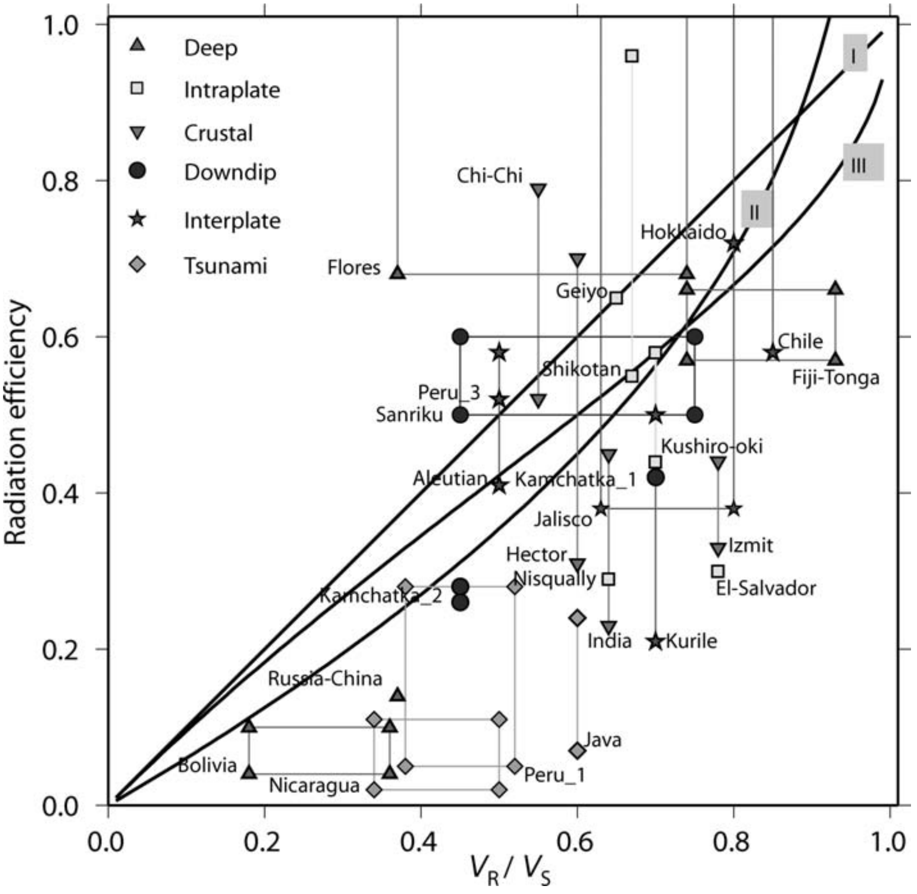


FIGURE 15.6 Radiation efficiency and rupture velocity. While for small earthquakes radiation efficiency is more or less constant, ~0.6, large earthquakes show the range given here. Symbols I, II, III refer to different crack models to describe earthquakes. (After Venkataraman and Kanamori, 2004.) Note: $c_L \equiv V_S$.

Table 15.1. An earthquake is approximated by two blocks moving relative to one another (Fig. 15.2), with total moment given by

$$M_0(t) = \frac{M_0}{2}(1 - \cos\omega t), \text{ for } 0 < \text{t} < \frac{\text{T}}{2}, \qquad (15.48)$$

where, taking into account the blocks on either side of the fault, $M_0 = 4\mu y_e L^2$. Using expressions from Table 15.1, Eq. (15.47) becomes

$$E_R = \frac{\omega^4 M_0^2}{4 \times 10\pi\rho V_S^5} \int_0^{\frac{T}{2}} \cos^2(\omega t)\mathrm{d}t = \frac{\omega^3 M_0^2}{80\rho V_S^5}. \qquad (15.49)$$

The total change in elastic energy is

$$E_{\text{total}} = \frac{\sigma_1 + \sigma_2}{2\mu} M_0, \qquad (15.50)$$

so that seismic efficiency, $\eta_S = E_R/E_{\text{total}}$ is

$$\begin{aligned}\eta_S &= \frac{\omega^3 M_0^2}{80\rho V_S^5} \frac{2\mu}{M_0(\sigma_1 + \sigma_2)} = \frac{\omega^3 \Delta\sigma l^3}{80\rho V_S^5} \frac{2\mu}{(\sigma_1 + \sigma_2)} \\ &= \frac{1}{40} \frac{\Delta\sigma}{(\sigma_1 + \sigma_2)}.\end{aligned} \qquad (15.51)$$

For the example chosen in Table 15.1, $\eta_S = 0.5\%$. Because we cannot normally measure σ_2, seismic efficiencies are uncertain. In cases where measurements have been made, efficiencies less than 6% have been inferred (McGarr, 1999). Radiation efficiency, Eq. (15.45), is obtained by replacing $(\sigma_1 + \sigma_2)$ in Eq. 15.51) by $\Delta\sigma$ (Eq. (15.39)):

$$\eta_R = \frac{1}{40}. \qquad (15.52)$$

Thus the block model gives $\eta_R = 2.5\%$ for all magnitudes. The low radiation efficiency is a consequence of the limited accelerations permitted by dynamical modelling of the motion of the sliding block. Kinematic models with assumed displacement-time histories that give higher accelerations allow much higher efficiencies. For example, Singh and Ordaz (1994) found that the Brune model (Eq. (14.32)) gives $\eta_R = 46\%$ and that if the model is modified to give a sharp corner frequency, $\eta_R = 86\%$. Ide *et al.* (2003) obtain a well-constrained value of 66% from borehole data.

There are a number of reasons why observed radiation efficiencies (Venkatamaran and Kanamori, 2004; Kanamori and Brodsky, 2004) are much higher than for the smooth sliding block model. Models of dynamically spreading cracks (Madariaga, 1976; Das, 1981) show that healing waves are generated at boundaries where the cracks stop. The healing waves quench the motion in the interior of the crack significantly earlier than friction alone. Another reason is the dependence of the coherence of high-frequency radiation on the spacing of near-field network stations. For coherence, the wavelength of the radiation must be longer than the station spacing (Dainty, 1995). This means that at shorter wavelengths the stations are receiving signals from different patches of fault and therefore the high frequency radiation is not generated by smooth motion of an extended fault, but by semi-independent motion of small patches. This can be attributed to variable friction and high local accelerations, with heterogeneous fracturing adjacent to the fault. This leads to an erratic moment–time history with a larger value of $\int \ddot{M}_0^2 \mathrm{d}t$ in Eq. (15.47).

If earthquakes have a seismic efficiency of 6%, as McGarr (1999) suggests, the question arises that why, if the remaining 94% of the energy is taken up as frictional heat, a heat flow anomaly is not observed on large strike-slip faults such as the San Andreas (Lachenbruch and Sass, 1980). The absence of a heat flow anomaly has prompted several speculations. The possibility that dynamic fault friction somehow becomes low has been considered, but this is difficult to reconcile with the high deviatoric stresses measured in boreholes (McGarr, 1999). Some energy is consumed in breaking rock and powdering it to produce fault gouge (see recent discussion by Abercrombie and Rice, 2005), but we find this to be inadequate, and a more plausible alternative is the flow of ground water as the principal mechanism for the removal of locally generated heat. A significant fraction of rainfall on land sinks in to become ground water. It cannot accumulate in the long term and flows to the sea by whatever channels are easiest, and fault zones are broken up and porous, making them easy paths. Flow of ground water extends much deeper than is often recognized, as emphasized by a comment on discoveries from the deep (KTB) borehole in Germany (Haak and Jones, 1997): ‘Another major surprise was the presence of abundant fluids . . . throughout the complete depth range . . . with the astonishing result that the formation pressure remains hydrostatic down to the base of the hole at 9100 m.’

15.7 Foreshocks and prediction ideas

The average interval between major earthquakes on sections of fault where they occur is typically 100 years or more. Thus, if we take the magnitude of the elastic strain release to be 2×10^{-4}, the average rate of its build-up between shocks is no more than $2 \times 10^{-4}/100\ \text{years} = 6 \times 10^{-14}\ \text{s}^{-1}$. We may compare this with the rate of change of tidal strain in the solid Earth. The strain amplitude of the lunar tide is about 5×10^{-8}, cycled in 12.4 hours and giving a maximum strain rate of $(2 \times \pi \times 5 \times 10^{-8}/12.4)\ \text{hour}^{-1} = 7 \times 10^{-12}\ \text{s}^{-1}$. This is 100 times the average rate of seismic strain build-up. But earthquake occurrences are not related to the tidal cycle. Therefore the concept of a sharply defined breaking stress appears irrelevant and the system is governed by time delays to instability (Eq. (15.18)).

The energy released during an earthquake is stored as elastic strain energy in the focal region immediately before the shock. Early efforts to find an earthquake prediction method were based on the supposition that this energy would be recognizable from measurements of stress, strain or related secondary effects. But, as noted in the preceding sections, seismic stress release does not normally exceed about 10 MPa (100 bar); this is less than the static stresses that are often found in seismically stable areas, so the expectation was rather that seismic stresses or strains would display a characteristic time-dependence, indicative of instability. In Section 15.3 we present a model of earthquake initiation, which depends on the instability of small patches that are probably too small to observe by prediction techniques currently under investigation. As suggested in Section 15.1, the hope may be in recognition of what we refer to as vulnerability rather than instability. If a given patch is to trigger others and grow into a large earthquake, it requires a general elevated vulnerability. Stress levels over a wide area would need to be high, approaching the failure stress. This has been termed critical instability, and it has been suggested that it may be detectable in a regional sense as long-range correlations of pre-event seismicity (Keilis-Borok, 2002), with accelerating foreshock moment release as critical patches are activated (Bowman and King, 2001).

Reasenberg (1999) notes that short-term earthquake clustering is the strongest non-random feature observed in earthquake catalogues. It allows one to obtain short-term probabilistic forecasts of future earthquake activity. No other phenomenon currently provides such a possibility. However, in that most large earthquakes are not preceded by identifiable foreshocks, they are, at present, unpredictable.

The Epidemic Type Earthquake Sequence (ETES) model, developed by Kagan and Knopoff (1987) and Ogata (1988), is based on the Gutenberg–Richter or Kagan laws (Section 14.7) to model the magnitude distribution, and Omori's law to characterize the decay of triggered seismicity after any shock. The model assumes that each earthquake may be simultaneously a mainshock, aftershock and/or foreshock. Up to a limit imposed by the corner magnitude of the frequency distribution represented by Eq. (14.44), each earthquake of magnitude M triggers aftershocks with a rate proportional to $10^{\alpha M}$, which decays with time according to Omori's law, $1/(t+c)^p$ (Eq. (15.27)). Thus each earthquake has a finite probability of triggering a larger one. The model gives the probability of triggered events as a function of space and time by superimposing the triggering probabilities of all preceding earthquakes.

The spatio-temporal distribution of triggered events at distance r and at time t after an earthquake of magnitude M is

$$\phi_M(r,t) = K \times 10^{\alpha M} \psi(t) f(r,M), \tag{15.53}$$

where K is a constant, α is the productivity of aftershocks from an event of magnitude M. $\psi(t)$is a normalized version of Omori's law,

$$\psi(t) = \frac{N}{(t+c)^p}, \tag{15.54}$$

and*f*(*r*,*M*) is a normalized distribution of horizontal distance between events. For example, Kagan and Jackson (2000) use functions of the form

$$f(r,M) = \frac{r}{\sigma_r^2} \exp[-r^2/(2\sigma_r^2)], \tag{15.55}$$

where σ_r is the scaling parameter (standard deviation) and increases with M. Equation (15.53) is fitted to historical catalogues, using statistical methods. The cumulative sum of probabilities has been used to construct probability maps against which future activity can be tested. They show clustering, of which the most obvious manifestation is aftershock sequences. The model, applied to catalogues from which recognized aftershock sequences have been removed, also identifies clustering (Kagan and Jackson, 1994), albeit weaker than for aftershocks. Major shocks around the Pacific rim are positively correlated in adjacent regions rather than occurring in seismic gaps, which was the earlier expectation.

16

Seismic wave propagation

16.1 Preamble

The consideration of elastic waves in this chapter is a preliminary to their use in Chapter 17 to study the internal structure of the Earth, but also an extension to the discussion of elasticity in Chapter 10. Virtually all the information that we have on the elasticity of the Earth is obtained from observations of seismic waves. Tidal deformation (Section 8.2) and the period of the Chandler wobble (Section 7.3) give supplementary data that extend to low frequencies, but lack the detail and precision of seismology. In the full context of the subject, seismological observations can be considered to extend to zero frequency if the static strains of earthquake displacements are included, but the lower limit of seismic wave frequencies is 3×10^{-4} Hz, the 54 minute period of the ${}_0S_2$ mode of free oscillation. At the other end of the scale, waves with frequencies of order 1 Hz are recorded at observatories remote from the earthquakes that generate them; for more local studies, and especially in exploration seismology, much higher frequencies are used. With this wide frequency range it is necessary to recognize a slight frequency dependence of elasticity and consequent dispersion of elastic waves (Section 10.7). Without an allowance for this there is a small, but noticeable, discrepancy between models of the Earth derived from high-frequency body waves (Section 16.2) and free oscillations (Section 16.6).

We recognize also a contribution to frequency dependence that arises from heterogeneity. Variations in elastic properties on a scale that is small compared with the wavelengths of seismic waves are averaged in the manner of the unrelaxed modulus considered in Section 10.4. Wavelengths that are shorter than the scale of heterogeneities introduce a local refraction/diffraction problem. A patch of low velocity material delays waves passing through it and may lose its effect if the faster waves in the surrounding medium diffract around it, causing the wavefronts to 'heal' and obscure the existence of the patch. Conversely, waves passing through a high velocity patch will be refracted so that they spread out, making the patch appear larger to an observer on the far side (Section 16.3). The effect is to bias the wave speed to a higher value than seen by longer waves that take a gross average of the elasticity. This was drawn to attention by Wielandt (1987) and is the subject of a detailed analysis for the case of random heterogeneities (Roth *et al.*, 1993). It gives a frequency dependence to seismic velocities additional to that arising from anelasticity and considered in Section 10.7.

It is convenient to distinguish three types of wave, body waves, surface waves and free oscillations, although there is no sharp distinction between them. High-frequency body waves (Sections 16.2–16.4) have wavelengths that are short compared with the curvature of the Earth and their propagation has a close analogy in the propagation of light waves in an optical system. The analogy leads to the concept of seismic rays, the normals to the wave fronts. These are refracted and reflected in the same way as light rays, but one difference must be noted. There is no optical equivalent to the compressional (P) waves, but only to the transverse (S) waves of seismology. As we consider waves of lower and lower frequencies, with correspondingly longer wavelengths, the body wave ray theory becomes less appropriate. For wavelengths that are significant fractions of the radius of the Earth we must consider instead the normal modes of the Earth as a whole (Section 16.6). These are the free oscillations, standing waves of long wavelength. Similarly, surface waves (Section 16.5) of long wavelength are better treated as free modes for which the sphericity of the Earth is properly taken into account. The free oscillations now have a prominent role in modelling the Earth. Synthetic seismograms are calculated as weighted sums of the normal modes, where the weights are determined by the elements of the moment tensor of an earthquake multiplied by the strains of the modes calculated at its location (Section 16.7). When modes are added in this way they simulate a total seismic record, which is seen to progress from identifiable P, S and surface waves through to reverberations of the Earth as a whole as the higher frequencies die away. Inversion of seismograms observed at stations of the global network using normal mode synthetics is now routine, and provides a catalogue of moment tensors for earthquakes worldwide.

16.2 Body waves

As the words imply, body waves are transmitted through the interior of the Earth. In a uniform body they would spread out as spherical wavefronts, but heterogeneities in the Earth refract and reflect them in numerous ways. The study of travel times of body waves to different distances led to our understanding of the Earth's internal structure (Chapter 17). They are easier to comprehend than the surface waves that are constrained to follow the surface and the layering close to it. The two kinds of body wave are referred to as P-waves and S-waves. P is simply the initial for primary, because P-waves arrive first. S (secondary) waves are slower. P-waves are compressional waves, involving alternating compressions and rarefactions of a medium, and are transmitted by liquids and gases as well as solids. The particle motion is in the direction of propagation. S-waves are shear waves, with particle motion transverse to the direction of propagation. They are transmitted only by solids. In the case of S-waves, different polarizations are distinguished, the plane of polarization being the plane containing the directions of propagation and particle motion. If the plane of polarization is vertical, the waves are designated SV, and if the particle motion is horizontal they are termed SH. These polarizations behave differently at horizontal boundaries.

The speed of a body wave in any medium is given by the square root of the ratio of an elastic modulus to the density. In each case the relevant modulus is the one appropriate to the material deformation. Definitions of elastic moduli and relationships between them are given in Section 10.2 and Appendix D. In an isotropic solid the P-wave speed is

$$V_P = \sqrt{\chi/\rho} = \sqrt{\left(K + \frac{4}{3}\mu\right)/\rho}, \qquad (16.1)$$

and the S-wave speed is

$$V_S = \sqrt{\mu/\rho}. \qquad (16.2)$$

Since S-waves cause pure shear of a transmitting medium, Eq. (16.2) is correspondingly simple, but the P-wave speed (Eq. (16.1)) is less obvious.

A sound wave in fluid causes alternating compressions and rarefactions that are hydrostatic, because the medium supports no shear stress. The modulus is therefore incompressibility (bulk modulus), K, and Eq. (16.1) applies with $\mu = 0$. Since compression in one direction involves material deformation in solids it is resisted by μ as well as by K. Consider first the transmission of a compressional wave along a rod that is thin compared with the wavelength of the wave. The compressions and rarefactions of the rod are then accompanied by transverse strains, so that each section dilates or contracts laterally by Poisson's ratio times the longitudinal compression or extension. The wave speed is $\sqrt{E/\rho}$, where E is Young's modulus for the material. This is the most commonly considered elastic modulus and is sometimes referred to by engineers as *the* elastic modulus. It may be measured for the material of a wire simply by stretching the wire. Now consider waves of progressively shorter wavelengths. When the wavelength is not much greater than the rod diameter the lateral contractions and dilations of adjacent sections (half-wavelengths) of the rod are no longer independent but oppose one another, and when the wavelength is very short compared with the diameter they are prevented from occurring at all. Then the modulus describing the axial compressions and dilations by the wave is the modulus of axial strain, χ, which is related to the other moduli by expressions in Appendix D. It exceeds Young's modulus, E, by a factor that depends on Poisson's ratio, ν, because axial strain becomes harder when the lateral response is prevented. In the Earth, wavelengths of P-waves are normally very short compared with the dimensions of the Earth or of its major layers and so the P-wave modulus is χ. Since there is a greater fundamental interest in the bulk modulus, K, it is convenient to replace χ by $(K + 4\mu/3)$, as in Eq. (16.1).

In an isotropic medium a seismic wave travels in a direction normal to the wavefronts, leading to the concept of seismic rays. This is useful in describing the propagation of body waves through internal layers or features of the Earth that are large compared with the wavelength. As in optics, ray theory is applicable only to situations in which wavelength is short compared with features of interest. Otherwise diffraction is important, requiring a wave theory.

Angles of refraction and reflection of seismic rays at material boundaries in the Earth are represented by equations that are exact analogues of the laws of optical reflection and refraction. With respect to the fraction of incident energy that is reflected or refracted, the seismic situation requires allowance for P to S and S to P conversions, which have no optical equivalent. But, as in optics, refraction occurs, however gradual the transition in wave speeds between media, whereas reflection, and wave conversion in the seismic case, diminish as the thickness of a transition becomes comparable to the wavelength.

Consider the refraction of a plane wave at a plane boundary, as in Fig. 16.1. Each point in a wavefront may be regarded as a source for further propagation of the wave (Huygens's principle) and the distance travelled in time Δt is proportional to the wave speed, which is different in the two media, as illustrated. The angles i_1 and i_2 are related by considering the triangles ABC and ABD. Thus

$$\sin i_1 = \mathrm{BC}/\mathrm{AB} = v_1 \Delta t/\mathrm{AB} \tag{16.3}$$

and

$$\sin i_2 = \mathrm{AD}/\mathrm{AB} = v_2 \Delta t/\mathrm{AB}, \tag{16.4}$$

so that

$$\frac{\sin i_1}{\sin i_2} = \frac{v_1}{v_2}. \tag{16.5}$$

This is Snell's law of refraction.

Similar constructions may be used for reflection and wave conversion. For a simple reflection there is no change in wave speed and so the angles of incidence and reflection are equal.

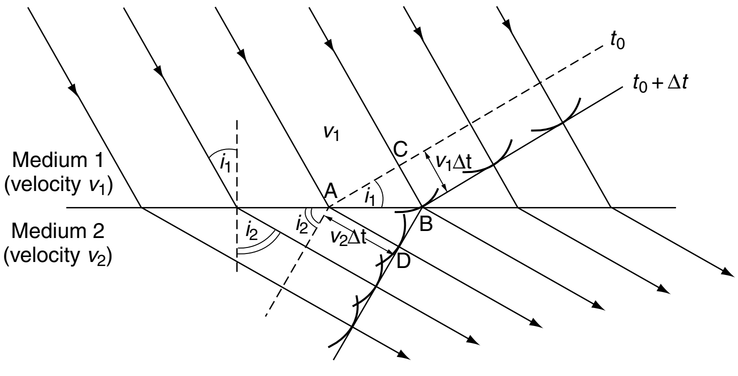


FIGURE 16.1 Huygens's construction for refraction of a wave at a plane boundary. Positions of a wavefront at times t_0 and $(t_0 + \Delta t)$ are shown by broken and solid lines and arrows on the rays show the direction of propagation. Each point on the wavefront at t_0 acts as a source of wavelets (shown as short arcs), the envelope of which is the wavefront at $(t_0 + \Delta t)$.

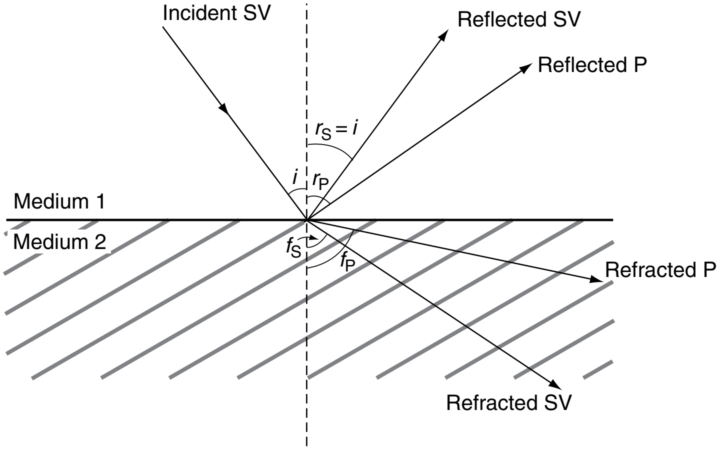


FIGURE 16.2 Reflected and refracted rays derived from an SV ray incident on a plane boundary. The boundary is assumed to be 'welded', that is, there is a continuity of solid material across it.

When partial conversion of a wave occurs, as illustrated in Fig. 16.2, then the speeds of the incident and converted waves, whether refracted or reflected, are used in Snell's law (Eq. (16.5)). For the wave speeds V_P (for P-waves) and V_S (for S-waves) in media 1 and 2, as in Fig. 16.1, the law becomes

$$\frac{\sin i}{V_{S1}} = \frac{\sin r_S}{V_{S1}} = \frac{\sin r_P}{V_{P1}} = \frac{\sin f_S}{V_{S2}} = \frac{\sin f_P}{V_{P2}} = p, \qquad (16.6)$$

where p is the ray parameter for the family of rays. Particular reflections or refractions cannot occur if this equation would require the sines of the relevant angles to exceed unity. Thus, a wave may be totally internally reflected in medium 1 if $\sin r_P$, $\sin f_S$ and $\sin f_P$ all exceed unity, although continuity of particle motion at the boundary requires participation by a layer of medium 2 with a thickness comparable to the wavelength. The inverse velocity $(1/V)$ is referred to as the slowness. Snell's law, Eq. (16.6), shows that the family of rays has a common horizontal component of slowness, given by p.

The incident wave considered in Fig. 16.2 is an SV-wave, a shear wave in which the plane of polarization or particle motion is vertical. This gives it a component of motion normal to a horizontal boundary. It is this component of the motion that generates P-waves at the boundary. An SH-wave, in which particle motion is parallel to the boundary, causes no compressions or rarefactions across the boundary and so can be refracted and reflected only as an SH-wave. Conversely, P-waves incident on a boundary generate both reflected and transmitted SV-waves, but no SH-waves. In general S-waves have both SV and SH components, but if a wave arrival at a seismic station is observed to be of pure SV type, then it is probably a conversion from a P-wave at an internal boundary.

Snell's law is a consequence of Fermat's principle that the travel time for a seismic wave between source and receiver along a ray path is stationary with respect to adjacent paths. In most cases the stationary point is a minimum. However, for reflection from a free surface it is a maximum relative to adjacent incident angles (Problem 16.2). In heterogeneous media the Fermat path is found by ray-tracing methods that vary the path until travel time is a minimum.

16.3 Attenuation and scattering

Plane seismic P-waves satisfy the equation of motion given in general form as Eq. (11.48). For propagation in one direction in an isotropic medium we obtain the familiar wave equation

$$\chi \frac{\partial^2 u}{\partial x^2} = \rho \frac{\partial^2 u}{\partial t^2}. \qquad (16.7)$$

This equates the force per unit volume (left-hand side) to the mass per unit volume times acceleration (right-hand side) and $\chi = \lambda + 2\mu = K + (4/3)\mu$

is the P-wave elastic modulus (Eqs. (16.1), (11.14), (11.16)). u is the displacement in the direction of propagation, and ρ is the density. Solutions to the wave equation can be written $u = f(x - ct)$ or $f(x + ct)$ depending on whether the wave is travelling in the positive or negative x-direction respectively, where c is speed. Differentiation and substitution in Eq. (16.7) gives $c = \sqrt{\chi/\rho}$. An arbitrary pulse can be represented as a sum of sinusoids (Fourier transform) but incorporation of attenuation or wave decay requires additional exponential factors. It is common practice in seismology to represent f in terms of complex exponentials, $\exp(\mathrm{i}\omega t)$, with both real and imaginary components. Particle displacement for a wave travelling in the positive x-direction can be written

$$u = A\exp[\mathrm{i}(kx - \omega t)], \tag{16.8}$$

where, if there is no attenuation, wave number is simply $k = \omega/c$. If the medium is attenuative, k is complex. It is convenient to retain k as the real component and add an imaginary term: $k + \mathrm{i}\alpha$. Then we can write

$$u = A\mathrm{e}^{-\alpha x}\exp[\mathrm{i}(kx - \omega t)], \tag{16.9}$$

which decays exponentially with distance, where α is known as the attenuation coefficient. This can be written in terms of the Q of the medium as $\alpha = \omega/2cQ$ (Eq. (10.21)).

As explained in Section 10.7, attenuation, that is non-zero α or $1/Q$, causes dispersion (except for the special case $Q \propto \omega$). This is a consequence of the requirement that a pulse should be causal, with no feature travelling faster than the wave speed. The general expression relating frequency dependences of wave speed and attenuation is the Hilbert transform in Eq. (10.30). When Q and k vary, ray-tracing methods are used to find the ray path, and attenuation effects are integrated along the ray.

Amplitude also decreases by geometric spreading. An expanding curved wavefront diminishes in amplitude as 1/(radius of curvature). Conversely, a contracting curved wavefront focusses energy. The amplitude variation is simple if the radius of curvature is very large compared with the wavelength, but if not, then a more general treatment is required to take account of diffraction. We adopt the solution from the equivalent optical problem (Born and Wolf, 1965, p. 441). The amount of focussing depends on the ratio of the aperture to the wavelength, λ, as

$$A(0)/A(r_0) = \frac{\pi R^2}{r_0\lambda}, \tag{16.10}$$

where $A(r)$ is amplitude at distance r from the focus, $2R$ is the aperture (chord) of a wavefront in the form of a spherical cap of radius of curvature r_0. As shown below (Eq. (16.14)), this is proportional to the ratio of the aperture area (πR^2) to the area of the inner Fresnel zone ($\pi r_0\lambda$), within which wave interference is constructive.

An essential feature of Eq. (16.10) is focusing that depends on wavelength. This is not related to dispersion and requires no frequency variation of seismic velocity, but is a consequence of diffraction. This phenomenon offers an explanation for concentrated patches of damaging intensity in Santa Monica, California, caused by the 1994 Northridge M 6.7 earthquake, which had an epicentre 21 km away, too distant to have caused the damage without strong focusing (Davis *et al.*, 2000). Santa Monica sits on a deep (~3 km) basin of sediment with an S-wave speed about half that of the basement rock. Irregularities in the boundary of the basin act as lenses, with dimensions (apertures) of order 1 km, focusing the arriving waves on patches determined by the direction of arrival. The focusing is frequency dependent, allowing the positions and dimensions of several such lenses to be calculated from Eq. (16.10).

Scattering of seismic waves from heterogeneities in an elastic medium causes amplitude decay additional to that arising from anelastic damping, which is discussed in Section 10.5. When a wave encounters elastic heterogeneities, energy is scattered in all directions, reducing the forward travelling energy. We identify separate scattering (s) and intrinsic (i) components of the attenuation coefficient:

$$\alpha = \alpha_\mathrm{s} + \alpha_\mathrm{i} \tag{16.11}$$

where 'intrinsic' implies a material property, that is local conversion of elastic energy to heat. Equivalently, the total Q is given by

$$\frac{1}{Q}=\frac{1}{Q_i}+\frac{1}{Q_s}. \qquad (16.12)$$

Analytical calculation of scattering coefficients is possible only for simple heterogeneities, such as a spherical contrast or a crack. Most analyses deal with single scatterers and sum the scattered fields from multiple scatterers spread throughout a homogeneous medium. This is adequate if the density of scatterers is low. However, the full problem involves multiple scattering, in which a single-scattered field interacts with other scatterers generating secondary fields. The secondary fields interact again, and so on in an infinite series, but because the amplitude of the scattered wave depends on the contrast in properties, which in the Earth is usually not very large, the series converges rapidly.

The amount of scattered energy depends on the impedance contrasts and the sizes of the scatterers relative to the wavelength of the seismic waves. Treatments of scattering appeal to methods of full wave theory rather than ray theory (Section 16.2). Ray theory effectively assumes that the frequency of the radiation is infinite. In reality, seismic wave pulses are composed of a range of frequencies and associated wavelengths. Ray theory is applicable if the length-scale of the region under investigation is large compared with the wavelengths, otherwise the different frequency components are scattered differently and finite frequency effects must be taken into account. The forward propagation of a finite frequency wave can be calculated by summing Huygens wavelets, as considered in determining Snell's law in Section 16.2. Consider a point A on the wavefront of a plane wave connected by the most direct ray to a point B at distance h in front of the wave. The amplitude at B depends on those wavelets that constructively interfere with the wavelet from A. Let C be a point on the wavefront a distance y from A. Wavelets from C will be increasingly out of phase with those from A as the distance AC (i.e., y) increases. The phase difference is given by

$$d\phi = 2\pi(\sqrt{h^2+y^2}-h)/\lambda \approx \pi y^2/h\lambda. \qquad (16.13)$$

Provided $d\phi \leq \pi$, the wavelets interfere constructively. This condition defines the first Fresnel zone as

$$y_1 \leq \sqrt{h\lambda}. \qquad (16.14)$$

The higher-order Fresnel zones correspond to 2λ, 3λ, ... path differences but cancel out (Sheriff and Geldart, 1982), so the amplitude and travel time at B are most affected by heterogeneities in the cone that includes B and the first Fresnel zone.

The effect of a heterogeneity at A depends on its size relative to the first Fresnel zone. If the size of the heterogeneity is $a \ll y_1$, it has negligible effect, whereas if $a \geq y_1$ it has maximum effect. For a plane wave that passes through a spherical heterogeneity a patch of the downstream wavefront becomes bowed inwards or outwards depending on whether the velocity in the heterogeneity is lower or higher than in the surroundings. As the wave progresses, amplitudes exhibit focusing or defocusing as the perturbed wavefront contracts or expands. Scattering effects that are seen in the near-field, $h \approx a$, may be appreciable, but at greater distances the Fresnel zone enlarges as $\sqrt{h}$, and the importance of the wavelets that pass though the scatterer diminishes compared with those from the remainder of the Fresnel zone. The disturbance heals in the far-field as the effects of the scatterer diminish. This process, called *diffractive healing*, occurs as wavelets from the part of the Fresnel zone that lies outside of the scatterer dominate the total. At large distances the plane wavefront is restored. Diffractive healing accounts for the difficulty in imaging hot spot plume stems (Chapter 12) deep in the mantle where they are small compared with $\sqrt{h\lambda}$. For example, if the seismic wavelength is 10 km, for observations at a distance $h = 4000$ km, the plumes need to have radii larger than 200 km to be observable. While images of much larger features such as plume heads are observable at these distances, plume stems smaller than $\sqrt{h\lambda}$ require near-field measurements. Such diffractive healing is important for seismic tomography of the crust and uppermost mantle, for which it is assumed that waves from distant earthquakes travelling through relatively homogeneous mantle have been healed to nearly plane waves.

If, instead of a propagating plane wave, we consider spherical waves travelling between a point source and a receiver, the Fresnel zones vary with distance along the ray path (Fig. 16.3(a)). Let the total path length be L, and consider a point at distance x along the ray with y measured at right angles to it. The phase difference between the Huygens wavelet that travels from the source at $x = 0$ via y to the receiver at L and the wavelet travelling directly along the ray path is

$$\begin{aligned} \mathrm{d}\phi &= \frac{2\pi}{\lambda}\left(\sqrt{x^2+y^2}+\sqrt{(L-x)^2+y^2}-L\right) \\ &\approx \frac{2\pi}{\lambda}\left(x\left(1+\frac{1}{2}\frac{y^2}{x^2}\right)+(L-x)\left(1-\frac{1}{2}\frac{y^2}{(L-x)^2}\right)-L\right) \\ &\approx \frac{\pi y^2}{\lambda}\left(\frac{1}{x}+\frac{1}{(L-x)}\right). \end{aligned} \qquad (16.15)$$

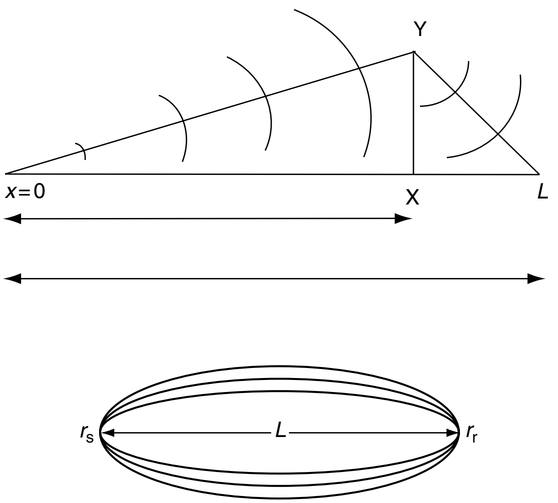


FIGURE 16.3 (a) Interference of Huygens wavelets. The phase difference between the direct wavelet and one that is diffracted from Y must be less than 180° for the wavelets to add constructively. (b) Fresnel volumes, as represented by Roth *et al.* (1993). These are zones in which direct and diffracted wavelets, travelling from the source (r_s) to the receiver r_r, interfere alternately constructively and destructively.

Again, for constructive interference $\mathrm{d}\phi < \pi$, so that

$$y < y_1 = [\lambda(L/x-1)]^{1/2}. \qquad (16.16)$$

The region traced out by $y_1(x)$, the *Fresnel volume*, is an ellipsoid-like volume about the ray (Fig. 16.3(b)). Wavelet contributions from the higher order Fresnel volumes (Fig. 16.3(b)) alternate between rapidly varying positive and negative interferences and so cancel out. Variations in properties within the inner Fresnel zone have greatest effect on both amplitude and phase of a seismic signal. Now we consider a typical seismic pulse. It is composed of a band of frequencies, each with a corresponding range of Fresnel zones. As a result, the perturbation of phase and amplitude is an integral of a sensitivity kernel $K(\mathbf{r})$ times the velocity perturbation field $v(\mathbf{r})$ over the volume, with maximum contributions from the inner Fresnel volumes corresponding to each wavelength (Spetzler and Snieder, 2004). Because the signal is band-limited, heterogeneities along the ray path that do not extend significantly into the Fresnel zone of the highest frequency component have negligible effect on the waveform. Fluctuations in this region are instead subject to diffractive healing. This gives rise to what has been termed the '*banana doughnut paradox*' by Marquering *et al.* (1999). The banana represents the ellipsoidal averaged Fresnel volume about a curved ray. The doughnut is a cross-section across the ray with hole smaller than the inner Fresnel zone associated with the highest frequency component. Thus, the ray itself traces out the locus of insensitivity to small-scale heterogeneities. Ray theory suggests the opposite; that changes lying along the ray should be observed at the receiver. In fact this assumption has been the basis for most tomographic interpretations. The paradox can be reconciled if one recognizes that ray theory used with finite- rather than infinite-frequency signals applies to heterogeneities much larger than the wavelength. To be detectable the heterogeneity must extend over the Fresnel volume either side of the ray, and this in turn is determined by the frequencies used.

From this discussion we see that objects that are small relative to the wavelength cause little scattering. As an extreme example, at the smallest scales seismic waves are not affected by crystal structures or grain distributions and the medium can be treated as homogeneous. Large objects can be treated by ray theory and analysed as piecewise homogeneous (Section 16.4). Scattering is most effective when the sizes of heterogeneities are comparable to that of the Fresnel zone. Seismic waves passing through a medium with random scatterers are most attenuated at frequencies for which associated Fresnel zones are equal to the mean size of the scatterers. For example, seismic body waves with frequencies of about 1 Hz are strongly scattered in the crust and upper mantle (Dainty, 1990; Padhy, 2005), giving rise to a band of attenuation at about 1 Hz. In a uniform, infinite, non-scattering medium, the S-wave pulse from an earthquake becomes zero after a time equal to the sum of rise time and propagation time of the rupture (Section 14.2), which is typically a few seconds for an earthquake of magnitude 6. However, reverberations, called coda, are observed to go on for many minutes. They are caused mainly by S–S scattering from heterogeneities (Aki, 1969; Zeng, 1993), and are thought to be comprised of back-scattered arrivals from an expanding hemi-ellipsoidal surface with foci at source and receiver. This ellipsoid maps the locus of scatterers for which travel times (source–scatterer–receiver) are equal for all scatterers along its boundary. As time progresses and the scattering ellipsoid grows proportionately, the coda decays by geometric spreading and intrinsic attenuation. Fresnel zones for 1 Hz S-waves in the crust ($\lambda \approx 3$ km) travelling typical distances of 30 km have a dimension of 10 km. That many geologic features, such as folded sediments and basins, are of this size may explain the 1 Hz ‘absorption’ band.

Recent advances in recording and processing digital seismic data from global and regional seismic networks have facilitated the study of scattering in the mantle and core. Haddon (1972) first suggested that precursive wave trains that build up before PKP body-wave phases were caused by scatterers in the mantle, and Haddon and Cleary (1974) concluded that they are concentrated in the D$''$ region. Vidale and Hedlin (2000) drew attention to sources of scattering near the core–mantle boundary (CMB) north of Tonga, requiring such large impedance contrasts that they have been presumed to include pockets of partial melt (but see an alternative interpretation in Sections 17.7 and 23.4). In addition to heterogeneities near the CMB there is a growing body of evidence for scatterers distributed throughout the mantle (Haddon *et al.*, 1977). Isolated scatterers observed in the mid-mantle may be delaminated slabs of subducted lithosphere (Kaneshima and Helffrich, 1999). Often the interpretation of the scattering in the mantle is not unique and a full description of the three-dimensional distribution of scatterers requires more information (Hedlin and Shearer, 2000; Earle and Shearer, 2001).

While the fluid outer core is observed to be homogeneous, scattered arrivals have been identified from the upper regions of the inner core (Vidale and Earle, 2000). The scattered signals from the inner core have been used to estimate inner core rotation. Using seismograms from Russian nuclear explosions recorded on an array in North America, Vidale *et al.* (2000) separated signals in the coda into those generated by scattering from the east and west sides of the inner core. They were able to show that the phase difference between east and west arrivals changed with time in a manner consistent with super-rotation of the inner core relative to the mantle. The inferred rate, $<0.2°$/yr, is consistent with the other estimates, as discussed in Sections 17.9 and 24.6.

Waves generated in a high-velocity slab tend to scatter out of the slab and so decay rapidly with distance. In contrast, waves generated in a low-velocity slab become trapped as angles of incidence at the boundaries become greater than the angle of critical internal reflection. Thus, it was surprising to find that earthquake waves generated in the subducting slab beneath Japan have higher amplitudes on the surface above the slab than expected from scattering, suggesting trapping of energy in the slabs. Furumura and Kennett (2005) showed that this can be explained by internal heterogeneities, such as high- and low-velocity laminations in the slab, that trap energy to form slab-guided waves. In a similar fashion scatterers in the lower crust, and possibly the uppermost mantle, trap energy from large surface explosions that travels as apparent P_n waves, refracted head waves (Section 17.2), at the crust–mantle boundary (Moho) to much greater distances than expected if the P_n were head waves generated beneath homogeneous crust overlying homogeneous mantle (Morozov and Smithson, 2000).

This discussion illustrates the importance of knowing the distribution of scatterers in the mantle when comparing Earth models derived from seismic waves of different frequencies. Roth *et al.* (1993) used finite difference calculations to examine travel time variations in a medium permeated with scatterers with higher than average velocity, and found that at high frequencies the wavefront is dominated by wavelets that have taken the fast paths through the high-velocity material. Thus, the travel time is shorter at high frequencies than that at low frequencies for which the waves respond to an average modulus of the medium. In Section 17.8 this phenomenon is mentioned as a possible reason for the discrepancy between models derived from body waves and normal modes.

16.4 Reflection and transmission coefficients at a plane boundary

The directions of waves refracted and reflected as plane waves are given by Snell's law (Eq. (16.6)). We are interested also in the relative amplitudes of reflected and refracted waves, the analysis of which gives some insight into the behaviour of guided waves. These include surface waves (Section 16.5) and also head waves that are the subject of Section 17.2. As we show, reflection and refraction amplitudes are controlled by the parameter known as acoustic impedance, which, for each medium and wave type, is the product of density and wave speed, ρV_P or ρV_S. Thus, observations of reflected waves give information about the contrasts in density and elasticity at boundaries. Observations of waves with different wavelengths may also show whether a boundary is sharp or marks a gradual transition in properties, because reflections occur only at boundaries that are sharp in the sense of having thickness much less than the wavelengths of waves used to observe them.

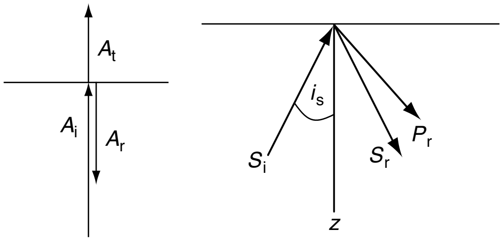


FIGURE 16.4 (a) Reflected and transmitted phases at normal incidence to an interface. (b) Reflected phases from a free surface at inclined incidence.

Consider a P-wave $u_z = A_i \exp[i(k_1 z - \omega t)]$, normally incident on a plane boundary at $z = 0$, separating materials of contrasting density and P-wave velocity given by, ρ_1, V_{P1} and ρ_2, V_{P2} (Fig. 16.4(a)) (the subscript i here indicates the incident wave and is not to be confused with $i = \sqrt{-1}$ in the exponential factor). Let the transmitted wave be $u_t = A_t \exp[i(k_2 z - \omega t)]$ and the reflected wave $u_r = A_r \exp[i(-k_1 z - \omega t)]$. Continuity of displacement requires

$$A_i + A_r = A_t. \tag{16.17}$$

Normal stress is given by

$$\sigma_{zz} = \chi \frac{\partial u_z}{\partial z} = ik\chi A e^{i(kz-\omega t)}. \tag{16.18}$$

Continuity of normal stress requires that

$$\begin{aligned} \sigma_{zz}^{i} + \sigma_{zz}^{r} &= \sigma_{zz}^{t}, \\ k_1\chi_1 A_i - k_1\chi_1 A_r &= k_2\chi_2 A_t. \end{aligned} \tag{16.19}$$

Let the reflection coefficient be R, transmission coefficient T and $A_i = A$.
From (16.17),

$$A + RA = TA \text{ or } 1 + R = T, \tag{16.20}$$

and from (16.19),

$$k_1\chi_1 - k_1 R\chi_1 = k_2 T\chi_2. \tag{16.21}$$

With $V_P = \sqrt{\chi/\rho}$, and $k = \omega/V_P$, Eq. (16.21) becomes

$$\rho_1 V_{P1} - \rho_1 V_{P1} R = \rho_2 V_{P2} T.$$

Combined with Eq. (16.20), this gives

$$R = \frac{\rho_1 V_{P1} - \rho_2 V_{P2}}{\rho_1 V_{P1} + \rho_2 V_{P2}}, \tag{16.22}$$

$$T = \frac{2\rho_1 V_{P1}}{\rho_1 V_{P1} + \rho_2 V_{P2}}. \tag{16.23}$$

These expressions illustrate the significance of acoustic impedance, ρV_P.

At a free surface Eq. (16.22) gives $R = 1$, $A_i = A_r$, so the reflected wave and the incident wave constructively interfere and the amplitude is doubled. Because the reflected wave adds constructively to the incident wave, the doubling of a pulse occurs over a depth range equal to half the dimension of the pulse. The stress at a free surface must be zero at all times. Note that the reflected wave is $u_r = A\exp[i(-k_1 z - \omega t)]$. So the stress is $\sigma_{zz} = -k_1\chi_1 A\exp[i(-k_1 z - \omega t)]$, whereas the incident wave had a stress given by $\sigma_{zz} = k_1\chi_1 A\exp[i(-k_1 x - \omega t)]$. The reflected and incident waves have opposite stresses, while the displacements are additive. At the surface, an incident tension is cancelled out by a reflected compression that travels downwards. The displacements have the same polarity relative to the z-axis, but relative to the direction of propagation, which for the reflected wave is the negative z-direction, the displacement is negative. Phase reversals of reflected waves occur whenever the impedance of the second medium is less than that of the first. Examples of 180° phase reversals include P-waves reflected from low velocity zones such as the core–mantle boundary, or the surface of the Earth.

At an interface with a medium of infinite acoustic impedance the amplitude would become zero. Then by Eq. (16.17) $A_i = -A_r$, that is the reflected and incident waves cancel at the surface, which is a node of a standing wave in the softer medium. Note that sometimes reflection and transmission coefficients are given for potentials instead of amplitudes, and we now use this convention.

Up to this point we have dealt with solutions to the one-dimensional wave equation in the form $\exp[i(kz - \omega t)]$. In considering reflection and refraction at boundaries that are not parallel to the wavefronts (Fig. 16.4(b)), the analysis of seismic displacements $\tilde{u}$ is simplified by decomposing it into gradients of (Helmholtz) potentials that allow P- and S-wave motions to be treated separately (e.g., Aki and Richards, 2002). The P-wave and S-wave components of $\tilde{u}$ are expressed as

$$\tilde{u} = \nabla\phi + \nabla \times \psi, \tag{16.24}$$

where the potentials ϕ and ψ satisfy P- and S-wave equations

$$\ddot{\phi} = V_P^2 \nabla^2 \phi, \tag{16.25}$$

$$\ddot{\psi} = V_S^2 \nabla^2 \psi. \tag{16.26}$$

In the two-dimensional case the P-wave equation becomes

$$\ddot{\phi} = V_P^2\left(\frac{\partial^2\phi}{\partial x^2} + \frac{\partial^2\phi}{\partial z^2}\right). \tag{16.27}$$

A solution to this equation is

$$\phi = \exp[i(k_x x + k_z z - \omega t)]. \tag{16.28}$$

Now we consider the problem of a wave travelling parallel to a boundary in the x-direction but with amplitude decaying in the z-direction. This corresponds to k_z being imaginary.

Let $k_z = i\omega\eta_P$.

$$\phi = \exp(-\omega\eta_P z)\exp[i(k_{Px} x - \omega t)]. \tag{16.29}$$

After differentiation and substitution in Eq. (16.27),

$$\frac{1}{V_{Px}^2} = \eta_P^2 + \frac{1}{V_P^2}, \tag{16.30}$$

where V_{Px} is the velocity in the x-direction. We see that $V_{Px} < V_P$ and depends on the value of η_P. Similarly for the S-wave potential, $\psi = \exp(-\omega\eta_S z)\exp[i(k_{Sx} x - \omega t)]$ and

$$\frac{1}{V_{Sx}^2} = \eta_S^2 + \frac{1}{V_S^2}. \tag{16.31}$$

Equations (16.30) and (16.31) describe inhomogeneous (P and S) waves that propagate in the x-direction and diminish with depth in the $\pm z$-direction. Consider a situation in which P-and S-waves have horizontal velocities equal ($V_{Px} = V_{Sx}$). Then by Eqs. (16.30) and (16.31) the depth dependences of the compressional and shear components are different ($\eta_P \neq \eta_S$). Such waves, travelling along a free surface with amplitudes that satisfy the free surface boundary condition, are Rayleigh waves (Section 16.5) with wave velocity $c_R = V_{Px} = V_{Sx}$, which, by Eq. (16.31), must be less than V_S because η_S is real.

At a surface or interface, the boundary conditions must be satisfied by matching the total stresses and displacements of all homogeneous and inhomogeneous waves. Because the η are squared they can be either positive or negative and still satisfy the wave equation. Such interface waves decrease or increase exponentially on either side of a boundary, but generally the increasing solution is discarded to avoid the unphysical situation of unbounded energy at infinity. Stoneley (1924) waves are P-SV waves (like Rayleigh waves) but at the interface of two half spaces (as discussed in detail by Aki and Richards, 2002, p. 156).

We now consider the case of a non-normal incident SV-wave at an angle i_S to a free surface (Fig. 16.2), at which there is no transmission. It will generate both a P-wave reflection at angle r_P and an S-wave reflection at angle i_S, as given by Snell's law (Eq. (16.6)), represented by the plane layer ray parameter $p = \dfrac{\sin r_P}{V_P} = \dfrac{\sin i_S}{V_S}$. (In the present discussion we are not concerned with spherical Earth geometry and it is convenient to use this version of the ray parameter, which we refer to as the plane layer ray parameter, to distinguish it from the parameter used with spherical Earth geometry and defined by Eq. (17.13).) In this case $k_x = \omega \dfrac{\sin r_P}{V_P} = \omega p, k_y = 0, k_z = -\omega \dfrac{\cos r_P}{V_P} = -\omega\xi$, where p and ξ are the horizontal and vertical slownesses (inverse speeds) for the P-waves and we set $\eta = \cos i_S / V_S$ to be the vertical slowness of the S-waves. In summary,

$$p = \frac{\sin r_P}{V_P} = \frac{\sin i_S}{V_S}, \tag{16.32}$$

$$\xi = \frac{\cos r_P}{V_P} = \sqrt{1/V_P^2 - p^2}, \tag{16.33}$$

$$\eta = \frac{\cos i_S}{V_S} = \sqrt{1/V_S^2 - p^2}. \tag{16.34}$$

Consider an incident S wave of unit amplitude given by

$$S_i = [u_x, u_z] = [\cos i_S, \sin i_S] \exp[i\omega(px - \eta z - t)]. \tag{16.35}$$

The reflected P-wave is

$$\begin{aligned} P_r &= [u_x, u_z] \\ &= R_{SP}[\sin r_P, \cos r_P] \exp[i\omega(px + \xi z - t)], \end{aligned} \tag{16.36}$$

and the reflected S-wave is

$$\begin{aligned} S_r &= [u_x, u_z] \\ &= R_{SS}[\cos i_S, -\sin i_S] \exp[i\omega(px + \eta z - t], \end{aligned} \tag{16.37}$$

where R_{SP} and R_{SS} are the amplitudes of the reflected P- and S-waves relative to the incident S-wave. R_{SP} and R_{SS} must satisfy the condition of zero traction on the free surface. We apply Eqs. (11.15) and (11.19) for the tractions, which are set to zero,

$$\begin{aligned} -2pV_PV_S\xi R_{SP} + (1 - 2V_S^2p^2)(1 - R_{SS}) = 0, \\ -(1 - 2V_S^2p^2)R_{SP} + 2V_S^3p\eta/V_P(1 + R_{SS}) = 0, \end{aligned} \tag{16.38}$$

yielding solutions

$$R_{SP} = \frac{4p\eta V_S\left(1/V_S^2 - 2p^2\right)/V_P}{R(p)}, \tag{16.39}$$

$$R_{SS} = \frac{\left(1/V_S^2 - 2p^2\right)^2 - 4p^2\xi\eta}{R(p)}, \tag{16.40}$$

where the common denominator is known as the Rayleigh function,

$$R(p) = \left(1/V_S^2 - 2p^2\right)^2 + 4p^2\xi\eta. \tag{16.41}$$

For the case of an incident P-wave, following the same procedure, it can be verified that $R_{PP} = -R_{SS}$ and $R_{PS} = R_{SP}(V_P^2\xi)/(V_S^2\eta)$. These coefficients are plotted in Figure 16.5(a). The reflected and transmitted waves show considerable variations in amplitude as functions of incidence angle. When $1/V_S > p > 1/V_P$ the reflected P-waves are inhomogeneous and the reflected S-wave is homogeneous (see Eqs. (16.33) and (16.34)). For an incident homogeneous P-wave

FIGURE 16.5(a) Homogeneous reflection coefficients for P-waves R_{PP}, R_{PS}, and incident S-waves R_{SS}, R_{SP}, from a free surface as a function of the plane layer ray parameter, p, with $V_P = 5$ km/s and $V_S = 3$ km/s.

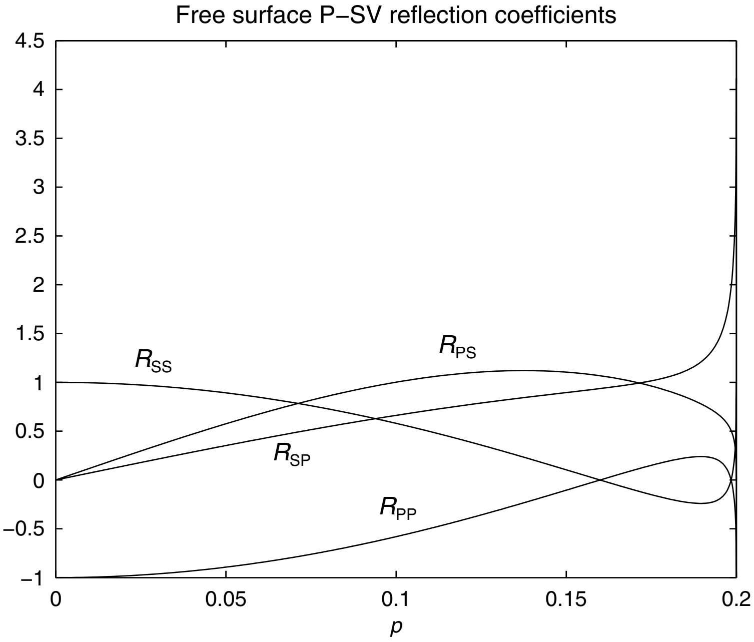
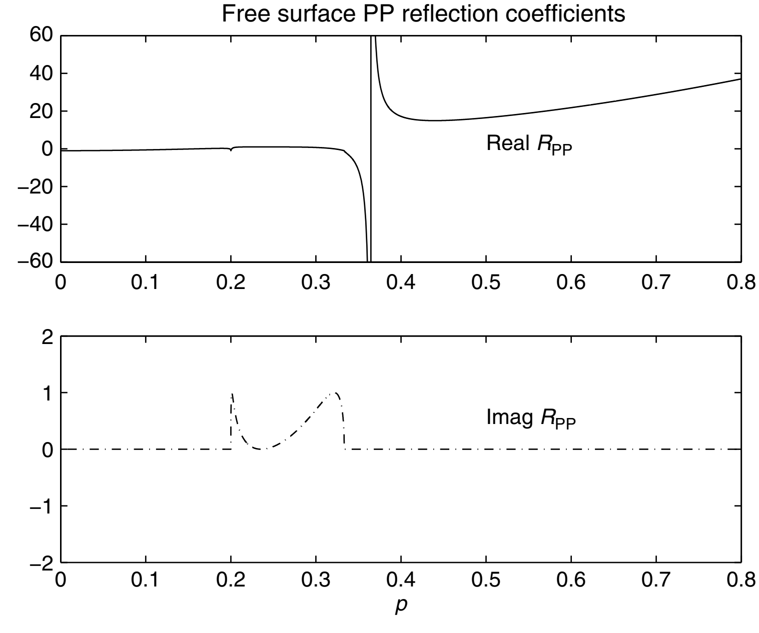


FIGURE 16.5(b) Real (solid curve) and imaginary (dotted) R_{PP} reflection coefficients as a function of the plane layer ray parameter, p (Eq. (16.32)), including the inhomogeneous range at $p > 0.2$, for the velocities assumed (as for Fig. 16.5(a)). The coefficient becomes singular at the Rayleigh pole, because $R(p)$ (Eq. (16.41)) vanishes. This value of p corresponds to the reciprocal of the Rayleigh wave speed, c_R, but gives a value slightly slower than observed because a very low value of V_P/V_S is assumed. The positive and negative infinities of R_{PP} at the Rayleigh pole are unphysical, but when the reflected P and S waves from a spherical source (obtained from R_{PP} and R_{PS}) are integrated (as in Eq. (16.48)) the result gives finite amplitude Rayleigh waves along the surface (Aki and Richards, 2002).



p cannot exceed $1/V_P$. If $p > 1/V_S > 1/V_P$ the interpretation is different. All waves along the interface, both incident and reflected, are inhomogeneous and R_{SP}, R_{SS} and p are related by Eq. (16.39) and Eq. (16.40) with complex angles since $\sin r_P = p\ V_P > 1$ and $\sin i_S = p\ V_S > 1$. Real and imaginary coefficients of R_{PP} are plotted in Fig. 16.5(b) for a range of p that includes inhomogeneous waves. Although we have considered a free surface, the analysis can be generalized to a boundary between contrasting media, with similar results. In general, at non-normal incidence an incident P- or S-wave generates four waves comprised of reflected and transmitted P- and S-waves. The four coefficients are referred to as the scattering matrix for the interface.

16.5 Surface waves

On seismic records of distant earthquakes (teleseisms), the waves of greatest amplitudes are generally surface waves that have followed the Earth's surface and not penetrated the interior. The exceptions are seismograms of deep focus earthquakes, which are not effective generators of surface waves, so that the body waves are more prominent. The dominance of surface waves on teleseismic records is due to the geometrical effect of wave spreading. Body waves spread out on wavefronts that are essentially spherical. Thus, the wave energy passing through any element of area diminishes as $1/r^2$, where r is the distance of travel from the focus. On the other hand, surface waves spread as an expanding circle across the surface. Thus, at near points the energy per unit length of wavefront falls off only as $1/r$, where r is now the radius of the circle. Moreover, this radius does not increase indefinitely but reaches a maximum when the waves have travelled 90° and beyond that it decreases again.

Body waves are almost non-dispersive and since wave energy density is proportional to the square of wave amplitude, body wave amplitudes diminish as $1/r$. The spreading of surface wave energy does not translate as directly into wave amplitudes, because surface waves are strongly dispersive. The waveform changes, becoming spread out in time, or equivalently, distance in the direction of travel. But, in spite of the dispersion, surface wave amplitudes decrease less with distance than do body wave amplitudes. In this circumstance it is convenient that body waves are faster than surface waves and so are not obscured by them on seismic records (Fig. 16.6).

There are two principal kinds of surface wave, both named after the originators of the theories describing them. Rayleigh waves appear as SV-waves, with a coupled P-wave component, as considered in Section 16.4, and Love waves propagate in the manner of SH-waves. In both

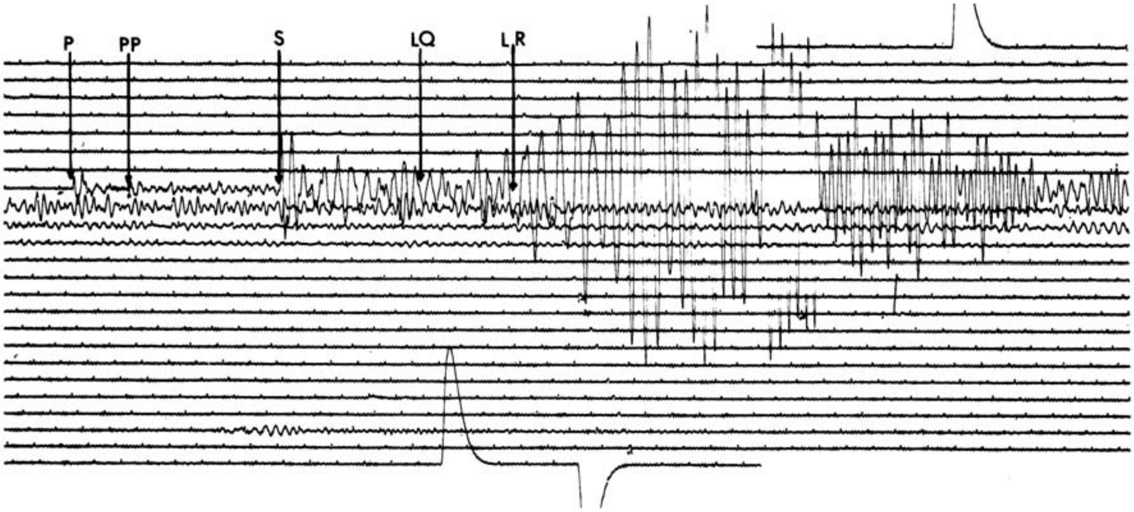


FIGURE 16.6 Seismogram obtained at Charters Towers, Queensland (station CTA), showing arrivals of P, PP, S and surface waves, LQ (Love) and LR (Rayleigh), from a magnitude 5.9 earthquake off northern Sumatra ($\Delta = 54^\circ.9$) on August 21, 1967. The successive lines are parts of a continuous helical trace and this figure shows about two thirds of the record that was unwrapped from the recording drum. There is one line per hour, with minute marks along the traces and calibration pulses at the beginning and end of the record. This recording is from a long-period east–west instrument, upwards on the record indicating eastward movement of the ground. The maximum amplitude of ground motion is about 200 μm.

cases the amplitudes of particle motion decrease away from the boundaries that guide the waves, becoming very small at depths greater than a wavelength. The propagation of surface waves has some similarity to the phenomenon of wave diffraction. The necessity for their existence is less intuitively obvious than is the existence of body waves, but there are analogies that help to give a feel for their behaviour.

Rayleigh waves resemble ocean waves. The material particles move in vertical ellipses which, at the surface, are retrograde, that is the direction of motion is that of a wheel rolling backwards relative to the progress of the wave. The restoring force is provided by the elasticity of the medium, and not by gravity as in ocean waves. Another difference is that the particle motion at depth is more complicated: the sense of the elliptical motion is prograde in the deeper parts, even for the fundamental mode, and there are higher modes with nodes in the motion. But, like open ocean waves, Rayleigh waves can propagate at the boundary of a uniform medium, or half space, no layered structure being required.

The Rayleigh wave speed, c_R, in a uniform half space is a slight function of its Poisson's ratio, ν, but for all reasonable values of ν, $c_R \approx 0.92V_S$ for the fundamental mode (see below, Eq. (16.45)), that is slightly less than the shear wave speed, V_S. In this simple case the waves are non-dispersive; longer wavelengths penetrate deeper but are simply scaled-up in size, and the ratio of restoring force to particle displacement, which determines the wave speed, is independent of scale. In a layered medium, such as the Earth, in which V_P, V_S increase with depth, longer wavelengths sample more of the higher-velocity material and in this circumstance Rayleigh waves are dispersive, the phase speeds of longer wavelengths being faster. It follows that observations of the dispersion give an indication of the velocity layering.

Analogies for Love waves are electrical waves in a wave-guide or light in a light pipe or optical fibre. They do not propagate on the boundary of a simple half space but require velocity layering with increasing shear wave speed, V_S, at depth. The simplest case is a uniform layer overlying a half space, which is also uniform, but with a higher value of V_S. Love waves behave like SH-waves that are made to follow the upper layer by repeated reflections, the angle of incidence to the boundary with the deeper material being greater than the critical angle for total internal reflection. Of course, since the two media are in welded contact, the motion at points along their common boundary is the same for particles on both sides of it. Thus, a Love wave propagates in both media, but the particle motion diminishes exponentially with depth in the lower one as an inhomogeneous wave. The penetration of the lower medium increases with wavelength, so the fraction of the wave energy propagating in the lower one increases with the ratio of wavelength to the thickness of the upper layer. Thus, Love waves are necessarily dispersive, with speeds, c_L, between those of shear waves in the two media. For the shortest wavelengths c_L is close to V_S of the upper medium and for the longest wavelengths it approaches the value of V_S in the lower one.

The eigenfunctions (solutions to the wave equation) for Rayleigh waves describe their disturbances in the vertical and horizontal directions as functions of depth (Fig. 16.7). They are combinations of exponential functions that decay with depth for both P and S motions.

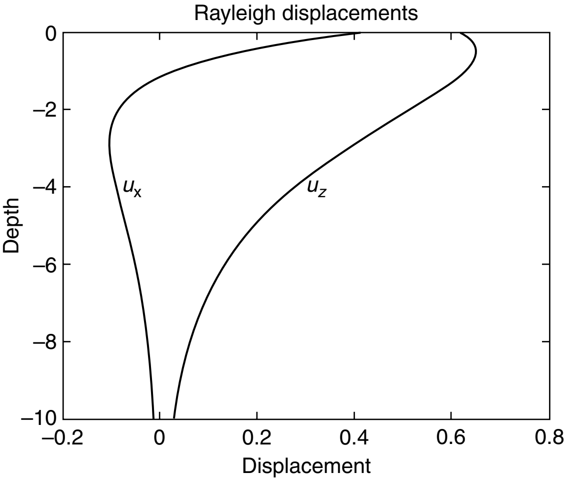


FIGURE 16.7 Horizontal and vertical displacement eigenfunctions for a fundamental mode Rayleigh wave as functions of depth below the free surface of a half space (normalized such that $A \sin i_P = 1$ m in Eq. (16.42)).

Many of the properties of Rayleigh waves can be deduced from elementary considerations (Knopoff, 2001) involving the stress-free boundary. The mathematical theory of surface wave propagation may be found in seismology texts, for example Aki and Richards (2002) and Bullen and Bolt (1985). Analytical treatments are tractable only in simple cases and, for the more complicated situation of the Earth, numerical methods are needed. The usual procedure is to adjust a model of the crust and upper mantle until it gives dispersion, the variation of wave speed with frequency, matching what is observed for a particular wave path. The wave speeds are sensitive to structure to depths of about a third of a wavelength. For Rayleigh waves we consider inhomogeneous P- and S-waves with the form of Eqs. (16.35) and (16.36) where ξ and η are imaginary (positive) and introduce amplitudes A and B for the P and S waves. The displacements are the real components of

$$\begin{aligned} P_{\mathrm{i}} &= \{u_x, u_z\} = A[\sin i_{\mathrm{P}}, \cos i_{\mathrm{P}}] \exp[\mathrm{i}\omega(px + \xi z - t)], \\ S_{\mathrm{i}} &= \{u_x, u_z\} = B[\cos i_{\mathrm{S}}, -\sin i_{\mathrm{S}}] \exp[\mathrm{i}\omega(px + \eta z - t], \end{aligned} \tag{16.42}$$

where the angles are imaginary. The stresses on the surface are equated to zero,

$$\begin{aligned} 2pV_{\mathrm{P}}V_{\mathrm{S}}\xi A + (1 - 2V_{\mathrm{S}}^2 p^2)B &= 0, \\ (1 - 2V_{\mathrm{S}}^2 p^2)A - 2V_{\mathrm{S}}^3 p\eta/V_{\mathrm{P}}B &= 0. \end{aligned} \tag{16.43}$$

In order that these homogeneous equations have a solution for A and B their determinant must equal zero,

$$\begin{vmatrix} 2pV_{\mathrm{P}}V_{\mathrm{S}}\xi & (1 - 2V_{\mathrm{S}}^2 p^2) \\ (1 - 2V_{\mathrm{S}}^2 p^2) & -2V_{\mathrm{S}}^3 p\eta/V_{\mathrm{P}} \end{vmatrix} = 0 = (1 - 2V_{\mathrm{S}}^2 p^2)^2 + 4V_{\mathrm{S}}^4 p^2 \xi\eta = R(p). \tag{16.44}$$

This is the Rayleigh function that arose in Eqs. (16.39), (16.40) and (16.41). After rearranging terms, Eq. (16.44) becomes a cubic equation in c_{R}^2,

$$\frac{c_{\mathrm{R}}^6}{V_{\mathrm{S}}^6} - 8\frac{c_{\mathrm{R}}^4}{V_{\mathrm{S}}^4} + c_{\mathrm{R}}^2\left(\frac{24}{V_{\mathrm{S}}^2} - \frac{16}{V_{\mathrm{P}}^2}\right) - 16\left(1 - \frac{V_{\mathrm{S}}^2}{V_{\mathrm{P}}^2}\right) = 0. \tag{16.45}$$

For a Poisson solid ($\lambda = \mu, \nu = 1/4$) the relevant real root of this equation is $c_{\mathrm{R}} = 0.92V_{\mathrm{S}}$. The Rayleigh wave is 8% slower than the S-wave. We can substitute back into Eqs. (16.42) and (16.43) to obtain the total normalized (by $A \sin i_{\mathrm{P}} = ApV_{\mathrm{P}} = 1\,\mathrm{m}$) horizontal and vertical displacements as functions of depth. Then the Rayleigh wave P and S components are given by

$$\begin{aligned} u_x &= \{\exp(-0.85kz) - 0.58\exp(-0.39kz)\} \\ &\quad \times \sin[k(x - c_{\mathrm{R}}t)], \\ u_z &= \{-0.85\exp(-0.85kz) + 1.47\exp(-0.39kz)\} \\ &\quad \times \cos[k(x - c_{\mathrm{R}}t)], \end{aligned} \tag{16.46}$$

where $k = \omega/c_{\mathrm{R}}$. These eigenfunctions are plotted in Fig. 16.7.

Rayleigh waves are generated if a point source, or distribution of point sources, is suddenly applied to an elastic half space. If the source is at depth the Rayleigh waves do not build up until the disturbance has travelled a distance such that the wavefront makes a sufficient angle with the surface to generate S. The deeper is the source the weaker is the disturbance by the time it reaches the surface; deep earthquakes are not effective generators of surface waves. A classical analysis in theoretical seismology is the response of an elastic half space to a point or line source, known as Lamb's problem (Lamb, 1904) with its extension to a buried source (Lapwood, 1949). Three main pulses are generated corresponding to the P, S and Rayleigh pulses. The mathematics is complicated, but exact solutions can be obtained using a method of inverting the integral transforms developed by Cagniard (1939, 1962) and DeHoop (1960), known as the 'Cagniard–DeHoop' method. A buried source generates P- and S-waves that convert and reflect at the free surface, subject to the zero traction boundary condition. Thus, as well as the incident waves, there are four reflected waves, giving six terms in the solution.

Consider a spherical P-wave, for example from an underground explosion, which is incident on a free surface. Let the source be buried at depth h below the free surface and a receiver a distance r away at depth z. At large distances a spherical wavefront approximates a plane wave governed by Snell's law, with $p = \sin i_{\mathrm{P}}/V$. At small distances, where the curvature of the wavefront is appreciable, it can be decomposed into a complex integral over the plane layer ray parameter p (Eq. (16.32)) of homogeneous and inhomogeneous plane waves (Aki and Richards, 2002).

$$P(\omega) \propto \int_{-\infty}^{\infty} \frac{\sqrt{p}}{\xi} \exp[\mathrm{i}\omega(pr + \xi z)\mathrm{d}p. \qquad (16.47)$$

For $|p| > 1/V_{\mathrm{P}}, \xi$ is imaginary and so by Eq. (16.33) this range in the integral corresponds to the inhomogeneous waves. The advantage of decomposing a spherical wave into plane waves is that the plane wave reflection and refraction coefficients can be used to model the effects of an interface on a spherical wavefront. In this case the reflected P- and S-waves are found by multiplying the integrand of Eq. (16.47) by R_{PP} and R_{PS} (see the argument following Eq. (16.41)). For example, the reflected P-wave is given by

$$P^{\mathrm{refl}}(\omega) \propto \int_{-\infty}^{\infty} R_{\mathrm{PP}} \frac{\sqrt{p}}{\xi} \exp[\mathrm{i}\omega(pr + \xi z + \xi h)]\mathrm{d}p \qquad (16.48)$$

(Aki and Richards 2002, Eq. 6.33). The integrand (16.48) becomes singular at values of p for which the denominator of R_{PP}, that is $R(p)$, becomes zero (Fig. 16.5(b)). This is known as the Rayleigh pole and, as we saw in Eqs. (16.44) and (16.45), is satisfied by a value of $p = p_{\mathrm{R}} = 1/c_{\mathrm{R}}$ slightly greater than $1/V_{\mathrm{S}}$ ($1/0.92V_{\mathrm{S}}$ for a Poisson solid).

The integral over p is equivalent to integrating plane (homogeneous and inhomogeneous) waves over all take-off angles from the source, weighted by R_{PP}. It can be solved numerically or approximately by the method of steepest descents (Aki and Richards, 2002) or, for a delta function source, by the Cagniard–DeHoop method. The integrand peaks for waves with a take-off angle given by Snell's law, $p_s = \sin i_s/V_{\mathrm{P}}$, (the saddle) and decreases exponentially (as a Gaussian) on either side. Another peak in the integrand occurs at the Rayleigh pole. For small angles of incidence, $p < p_{\mathrm{R}}$, we obtain the ray theory result because the exponential decay on either side of the saddle eliminates the contribution of the Rayleigh pole. However, as the incidence angle increases, p_s approaches p_{R}, and the contribution from the Rayleigh pole becomes more significant than the exponential decay. This occurs at an angle of incidence of about 80°. The reflected wave then contains, in addition to the reflected homogeneous wave, an inhomogeneous P wave, which is the P-wave component of the Rayleigh wave. A similar integral using R_{PS} has an identical pole, and generates the S-wave component of Rayleigh waves. We see that the coupled inhomogeneous P- and S-waves making up Rayleigh waves are generated by spherical waves interacting with a free surface. Above a buried source they build up beyond a distance about six times the depth.

In considering the dispersion of surface waves, it is essential to know whether it is the phase or group speeds of the waves that are observed. Normally, group speed, u, the speed of the envelope of a wave packet and therefore of wave energy, is observed, but in some cases it is possible to identify the speed v of a particular phase or feature of a wave. The relationship between group and phase speeds for a wave of wavelength λ and wave-number $k = 2\pi/\lambda$ is

$$u = v + k\frac{\mathrm{d}v}{\mathrm{d}k} = v - \lambda\frac{\mathrm{d}v}{\mathrm{d}\lambda} = -\lambda^2\frac{\mathrm{d}f}{\mathrm{d}\lambda}. \qquad (16.49)$$

In terms of frequency, f,

$$u = v\bigg/\left(1 - \frac{f}{v}\frac{\mathrm{d}v}{\mathrm{d}f}\right) \qquad (16.50)$$

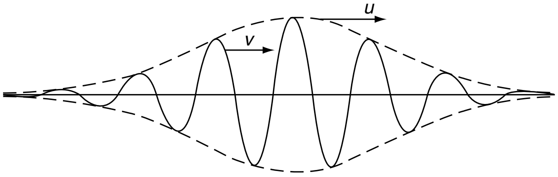


FIGURE 16.8 Phase and group speeds of a wave. The wave is represented by the solid line, with features travelling at the phase speed, v. The amplitude or envelope is outlined by the broken line which travels at the group speed, u. In physical situations such as the propagation of Rayleigh and Love waves, $u < v$, so that, as the individual waves or features pass through the envelope, they grow at the tail end and disappear out of the head or advancing front of the envelope. The energy of the wave travels at the group speed, u.

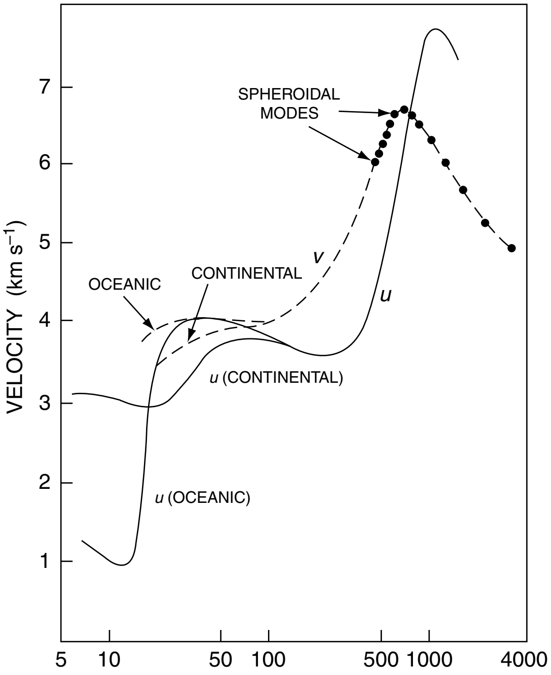


FIGURE 16.9(a) Fundamental mode Rayleigh wave dispersion. Group velocity, *u*, is shown by a solid line and inferred phase velocity, *v*, by a broken line, with the dispersion curve from free oscillation periods above 400 s. Figure based on Oliver (1962).

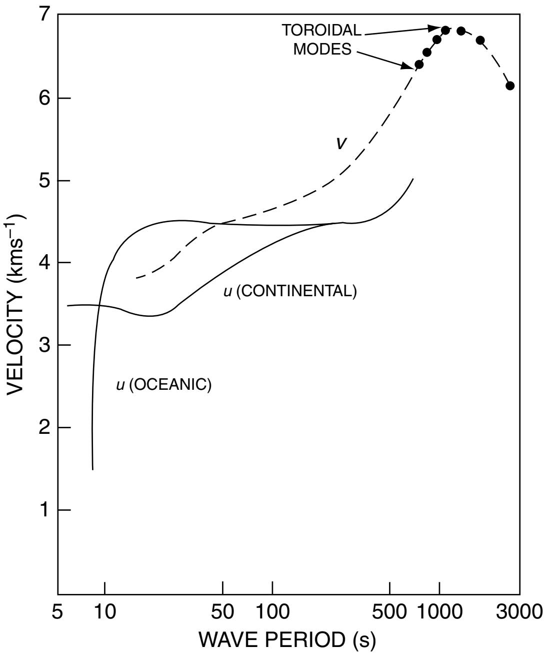


FIGURE 16.9(b) Dispersion curves for first mode Love waves. Group velocity, *u*, is shown as solid lines for continental and oceanic paths. Inferred phase velocity for continental paths is shown as a broken line, with free oscillation data above 750 s. Figure based on Oliver (1962).

or

$$\frac{1}{u} = \frac{1}{v} + f\frac{d}{df}\left(\frac{1}{v}\right) = \frac{1}{v}\left(1 - \frac{d\ln v}{d\ln f}\right) \\ = \frac{d}{df}\left(\frac{f}{v}\right) = \frac{d}{df}\left(\frac{1}{\lambda}\right). \tag{16.51}$$

The theory of surface wave propagation in media of specified elasticities and layer thicknesses gives the phase speed v from which u must be calculated by one of these equations. Over most of the range, $dv/d\lambda$ is positive for both Rayleigh and Love waves in the Earth, so by Eq. (16.49) $u < v$. This means that, as represented in Fig. 16.8, waves advance through the envelope representing their amplitude.

There are minima in the group velocity curves for both Rayleigh and Love waves, usually more noticeable for Rayleigh waves (Fig. 16.9). Thus, the last arriving waves of a wave train may be observed as a nearly sinusoidal wave of period corresponding to the velocity minimum. It has been called the Airy phase.

Excellent seismic records are required for the determination of surface wave dispersion. They must be spectrally analysed to find the speeds of individual frequency components of complex waveforms. At the longer periods, information derived from surface waves merges with that obtained from the periods of higher modes of free oscillation.

16.6 Free oscillations

The analysis of surface waves in terms of a flat Earth model is adequate at short periods, but becomes progressively less satisfactory at longer periods. Curvature of the Earth influences the dispersion and must be taken into account. A more general approach is to consider the modes of free oscillation. After a large earthquake the Earth resonates at numerous discrete mode frequencies, each of which may be thought of as a standing wave resulting from the superposition of oppositely travelling surface waves (or, in some cases, multiply reflected body waves). The resonant frequencies are related to the phase speeds of the corresponding surface waves, which are self-selected so that an integral number of wavelengths fits into the Earth. They therefore give a global average of the same information as the surface waves. Frequencies of more than 550 modes of free oscillation have been identified in seismic records (Masters and Widmer, 1995), providing a data set for Earth model studies that is independent of body wave travel times.

Many of the familiar names of classical physics and mathematics contributed to the theory of the vibrations of a sphere, notably S. D. Poisson, Lord Kelvin, H. Lamb, A. E. H. Love and Lord Rayleigh. Thus, it has been recognized for more than a century and a half that the Earth must have free modes, but for most of that time there was little expectation of observing them. There was also a problem that precise calculations of numerous mode periods for realistic Earth models would have been quite forbidding without electronic computers.

Interest in the subject was renewed in the 1950s by H. Benioff's development of an instrument to observe very long-period seismic waves. Most seismometers are inertial instruments, involving suspended masses that can be made sensitive to very long-period waves only with great difficulty. Benioff's instrument was a strain seismometer, a long quartz tube suspended in a tunnel, with one end fixed to the ground and the other attached to a displacement sensor to detect strain of the ground relative to the unstrained quartz. Such an instrument has a sensitivity to ground strain that is independent of wave period, subject only to a high frequency limit imposed by mechanical resonances in the mounting and by electronic response times.

From an examination of a record obtained immediately after a major earthquake in Kamchatka in 1952, Benioff tentatively identified an oscillation with an approximately 57 minute period as a fundamental mode of free oscillation. His report stimulated both instrumental and theoretical developments, so that, when the next really great earthquake occurred, in Chile in May 1960, several seismological research groups were able to record the oscillations that followed. The Benioff strain meter at the California Institute of Technology was most sensitive to toroidal oscillations, while across town at the University of California Los Angeles a tidal gravity meter (Slichter, 1967) recorded spheroidal oscillations but was insensitive to toroidal modes, for which there is no radial motion. Meanwhile Alterman *et al.* (1959) had calculated the frequencies of three modes, and some of their overtones, for several realistic earth models. Later in 1960, representatives of several groups met at an International Union of Geodesy and Geophysics (IUGG) meeting in Helsinki. Their observations agreed both with one another and with theory so well that a new branch of seismology was established on the basis of records of a single earthquake. Some of the mode frequencies were seen to have fine structure, a splitting of spectral lines due to rotation, ellipticity and heterogeneity of the Earth. The similarity to optical spectroscopy, in which spectral lines may be split by a magnetic field (the Zeeman effect), led to use of the expression *terrestrial spectroscopy* for free oscillation studies.

Another great earthquake occurred in Alaska in 1964 and the free oscillation records from both events (Fig. 16.10) were used to develop the first of a new generation of Earth models. Continued improvements in instrumentation, in data analysis and in methods of interpretation led to identification of a very large number of modes and to the use of records from smaller earthquakes. Evidence of continuous excitation of fundamental spheroidal modes was first reported by Nawa *et al.* (1998), on the basis of a superconducting gravity meter record from a quiet station in Antarctica and promptly confirmed by several other groups. In this case oceanic–atmospheric excitation is indicated (Rhie and Romanowicz, 2004). The use of free mode periods in model studies in the late 1960s and 1970s became so effective that little further improvement in our knowledge of the broad scale Earth structure can now be expected. The appearance in 1981 of PREM (Appendix F), the most widely used global model since that time,

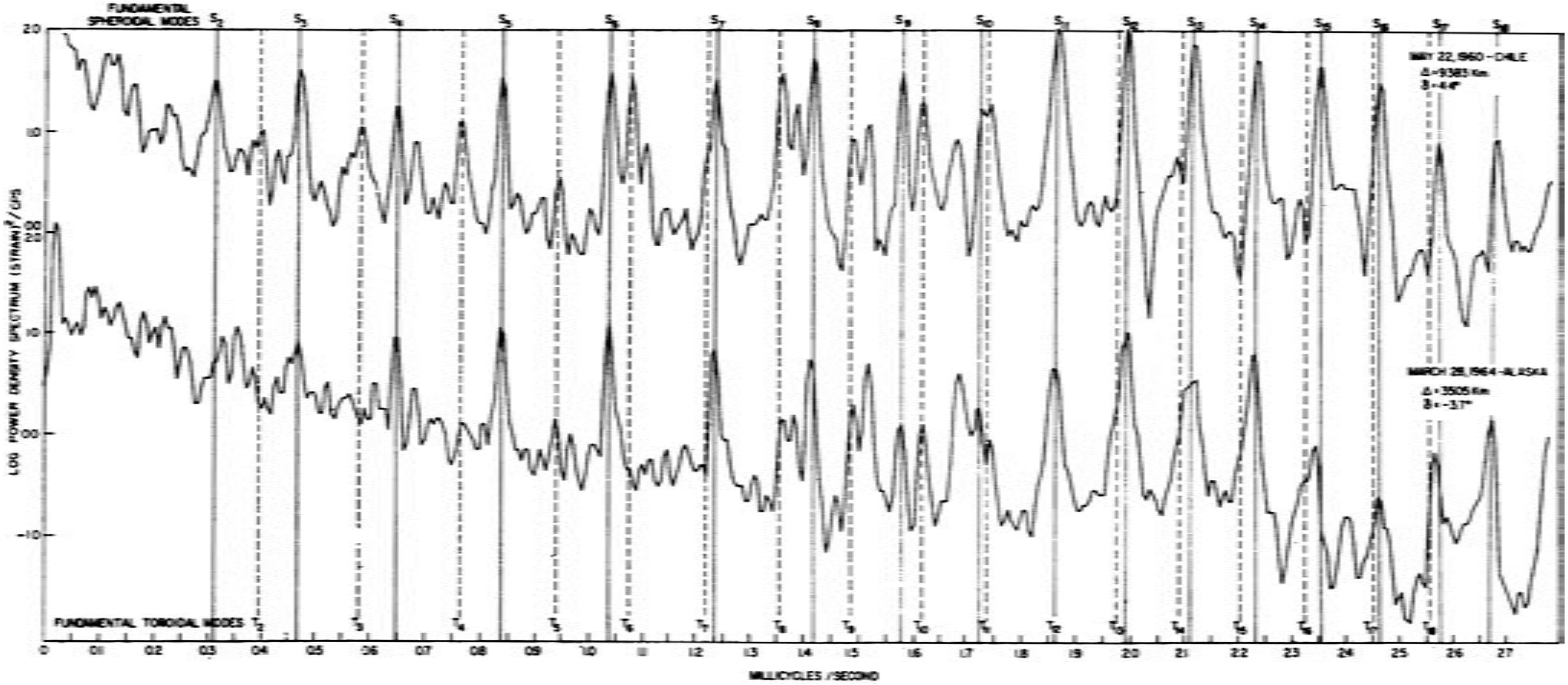


FIGURE 16.10 Spectra of Earth strain recorded at Isabella, California, for the Chilean 1960 and Alaskan 1964 earthquakes. δ is the angle between the strain seismometer axis and the great circle path to the epicentre. Reproduced, by permission, from Smith (1967).

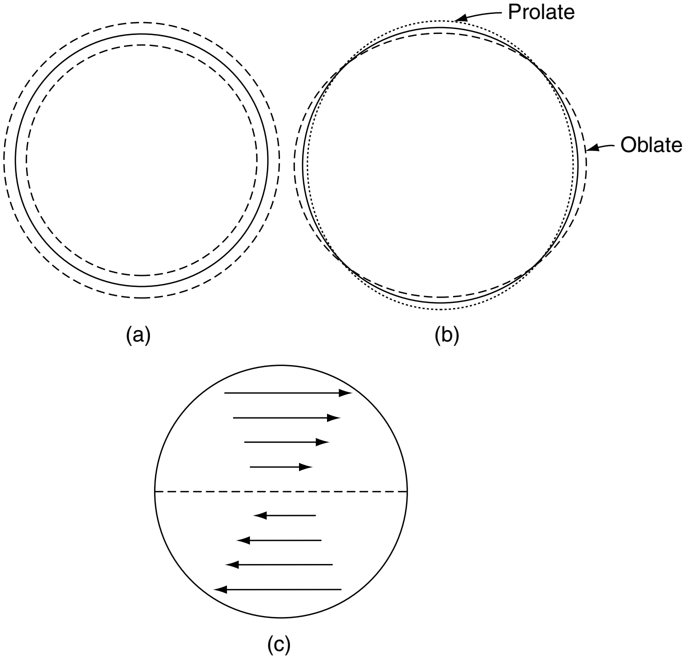


FIGURE 16.11 The simplest representative modes of free oscillation. (a) Radial oscillation$_0S_0$. (b) Spheroidal 'football' mode $_0S_2$. (c) Instantaneous angular motion in toroidal mode $_0T_2$.

signalled a change in scientific direction. At least in the mantle, lateral variations in structure now attract more attention than further refinements of the spherically averaged model, although equation of state studies (Chapter 18) have reached the point that an improved model is needed.

As mentioned above, there are two fundamentally different kinds of oscillation, spheroidal (S) and toroidal (T). In both cases the patterns of surface deformation take the form of spherical harmonic functions (Appendix C), which appear in solutions of the seismic wave equation in spherical geometry (Eq. (C.3)). The simplest to envisage are the radial modes, a special case of spheroidal oscillations in which the motion is purely radial. They are alternating dilations and compressions of the whole Earth (Fig. 16.11(a)), designated $_0S_0, {_1S_0}, {_2S_0} \dots$ with spherical nodal surfaces within the Earth, the instantaneous motion being opposite in adjacent layers. The prefix gives the number of these nodal surfaces and the following subscript is the number of nodal lines on the surface, that is the harmonic degree, l, of the surface pattern.

The slowest mode is $_0S_2$ (Fig. 16.11(b)), with a 54-minute period. This has been called the football mode, because the ellipsoidal deformation of the Earth alternates between prolate and oblate. There is no $_0S_1$ mode, because that would imply oscillation of the whole Earth mass, but $_1S_1$ occurs, the internal motion being opposite to that of the outer shell. An infinite series of fundamental spheroidal modes $_0S_l$ each has an infinite series of overtones $_nS_l$. The general case, $_nS_l^m$, involves a tesseral harmonic deformation of the surface in the form $P_l^m(\cos\theta)\cos m\lambda$ (see Appendix C), but the frequencies are almost the same for different values of m with the same l and n. They would be identical for a non-rotating, spherically layered Earth, but this degeneracy is broken by rotation, causing the line splitting mentioned earlier. Ellipticity and lateral heterogeneity give smaller contributions to the line splitting. The splitting decreases with mode frequency and is important only for the low degree modes. The superscript m is generally omitted in referring to free oscillation modes.

Of the toroidal modes, $_0T_2$ (Fig. 16.11(c)) is the simplest, being an alternating twist between two hemispheres. Higher modes involve the subdivision of the surface motion into 3, 4, ..., l zones and, as with spheroidal oscillations, there are overtones with internal nodal surfaces, whose number is given by the prefix n. As in the case of spheroidal oscillations, modes represented by tesseral harmonics have almost the same frequencies as zonal harmonics of the same degrees.

The spheroidal modes are standing Rayleigh waves and the toroidal modes are standing Love waves in spherical geometry. A particular advantage in analysing surface waves by free mode theory is that the sphericity of the Earth is properly accounted for. Clearly this is more important for the longer wavelengths and is essential at the longest wave periods in Fig. 16.9.

The surface patterns of free oscillation motions are spherical harmonics, as mentioned above. However, the coordinate axes are not the geographic axes but are controlled by the locations and fault orientations of the exciting earthquakes. Thus, any particular observatory lies close to some nodal lines for one earthquake, but to a different set after another earthquake. When this geometrical variation is allowed for, estimation of the relative excitation amplitudes of the various modes gives information about the earthquake mechanisms. For this purpose the amplitudes should be extrapolated back to a time immediately after an earthquake as there is a westward drift of the nodal pattern that can confuse amplitude observations for the lower modes. This drift must be recognized also in using amplitude decay as a measure of damping. The nodal drift is related to the mode splitting and arises from the slightly different phase speeds of eastward and westward propagating waves in the rotating Earth.

Early Earth models, developed by K. E. Bullen, used the seismic wave velocities, V_P and V_S, constraining the density structure to match the total mass and moment of inertia. Since the 1960s, Earth models have increasingly relied on the frequencies of the modes of free oscillation. The torsional modes involve only horizontal motion, but the spheroidal modes involve changes in shape, with radial motion and therefore gravitational as well as elastic restoring forces. This means that they provide an independent fix on the density structure. We can understand this in terms of a simplified model that is amenable to elementary calculus, a homogeneous sphere subject to self-gravitation, with density and elasticity uniform throughout, ignoring the effect of the steady internal compression.

Consider the spheroidal mode ${}_0S_2$, an oscillation with alternating prolate and oblate ellipsoidal deformations. Using cylindrical coordinates, with the origin at the centre and axes z along the symmetry axis, r in the perpendicular plane and θ the angle to the z axis, the surface deformation has the form of the zonal harmonic $P_2(\cos\theta)$ (Appendix C) and is given by

$$\delta = aP_2 \sin\omega t = a(3/2\cos^2\theta - 1/2)\sin\omega t, \quad (16.52)$$

for which the maximum axial elongation is $\pm a$, with corresponding maximum radial contraction $\mp a/2$. We calculate the frequency of the oscillation, ω, by using the fact that the total energy is conserved, being entirely kinetic energy, E_K, at the instants of zero deformation and entirely potential energy of elastic strain, E_S, and displacement of mass in the gravity field, E_G, at the instants of maximum deformation. Thus

$$E_K = E_S + E_G. \quad (16.53)$$

E_G has a simple form. The energy of elevation by distance δ of any surface element dA, relative to the undeformed radius, R, is $(1/2)\rho g\delta^2 \mathrm{d}A$ because mass $\rho\delta \mathrm{d}A$ is elevated by an average distance $\delta/2$ against gravity g. Areas for which δ is negative are included by noting that the missing mass below R can be considered raised to R and then transferred to the areas of positive δ. Thus, with δ given by Eq. (16.52), at $\sin\omega t = 1$,

$$\begin{aligned} E_G &= \frac{1}{2}\rho g \int \delta^2 \mathrm{d}A = \frac{1}{2}\rho g a^2 \\ &\quad \int_0^\pi \left(\frac{3}{2}\cos^2\theta - \frac{1}{2}\right) 2\pi R^2 \sin\theta \, \mathrm{d}\theta \\ &= \frac{2\pi}{5} R^2 \rho g a^2. \end{aligned} \quad (16.54)$$

With substitution for g by $g = GM/R^2 = (4/3)\pi G\rho R$,

$$E_G = \frac{8\pi^2}{15} G R^3 \rho^2 a^2. \quad (16.55)$$

At maximum deformation the strain can be represented as two perpendicular shear strains, each with the form of Fig. 16.11(b). Each component of strain gives an extension by $a/2$ in the axial direction and $-a/2$ in the equatorial plane with the equatorial strains at right angles to one another, so that there is a uniform contraction $-a/2$ in all directions in the equatorial plane and total extension a of the axis. Thus, we consider two perpendicular shear strains, each of magnitude $\varepsilon = a/R$, that are uniform throughout the volume V of the sphere, and the strain energy is

$$E_S = 2\left(\frac{1}{2}\mu\varepsilon^2\right)V = \frac{4}{3}\pi R\mu a^2. \qquad (16.56)$$

In calculating the kinetic energy, E_K, we resolve the motion into axial (z) and radial (r) directions, in which the amplitudes of the motion are $az/R = a\cos\theta$ and $(1/2)ar/R = (1/2)a\sin\theta$ respectively. There is no radial motion on the axis and no axial motion in the equatorial plane. For sinusoidal motion at angular frequency ω, the maximum particle velocities in the axial and radial directions are $a\omega\cos\theta$ and $(1/2)a\omega\sin\theta$, and since they are mutually perpendicular, their contributions to the energy are simply added. For an elementary volume $\mathrm{d}V$ the energies are $\mathrm{d}E_z = (1/2)(a\omega\cos\theta)^2\rho\,\mathrm{d}V$ and $\mathrm{d}E_r = (1/8)(a\omega\sin\theta)^2\rho\,\mathrm{d}V$. To integrate the axial component we divide the sphere into elementary discs of radius $R\sin\theta$ and thickness $\mathrm{d}z = R\sin\theta\,\mathrm{d}\theta$ at $z = R\cos\theta$, so that $\mathrm{d}V = \pi R^3\sin^3\theta\,\mathrm{d}\theta$ and

$$E_z = \int_0^\pi \frac{1}{2}(a\omega\cos\theta)^2\rho\pi R^3\sin^3\theta\,\mathrm{d}\theta = \frac{2\pi}{15}R^3\rho\omega^2a^2. \qquad (16.57)$$

To integrate $\mathrm{d}E_r$ we divide the sphere into elementary cylinders of thickness $\mathrm{d}r = R\cos\theta\,\mathrm{d}\theta$ and length $2z = 2R\cos\theta$, so that $\mathrm{d}V = 4\pi R^3\cos^2\theta\sin\theta\,\mathrm{d}\theta$, giving

$$E_r = \int_0^{\pi/2} \frac{1}{8}(a\omega\sin\theta)^2\rho 4\pi R^3\cos^2\theta\sin\theta\,\mathrm{d}\theta = \frac{15}{\pi}R^3\rho\omega^2a^2, \qquad (16.58)$$

and therefore

$$E_K = E_z + E_r = (\pi/5)R^3\rho\omega^2a^2. \qquad (16.59)$$

Substituting for E_G, E_S and E_K by Eqs. (16.55), (16.56) and (16.59) in Eq. (16.53), we have

$$\omega^2 = \frac{8\pi}{3}G\rho + \frac{20}{3}\frac{\mu}{\rho R^2}. \qquad (16.60)$$

Recognizing that the shear wave speed is $V_S = (\mu/\rho)^{1/2}$ and that the travel time for a shear wave across a diameter is $T = 2R/V_S$, we can write the period, $\tau = 2\pi/\omega$, of the ${}_0S_2$ mode for a uniform sphere as

$$\tau = \left(\frac{2}{3\pi}G\rho + \frac{20}{3\pi^2}\frac{1}{T^2}\right)^{-1/2}. \qquad (16.61)$$

In the absence of the gravity term this would be $1.217T$, that is the mode period would be comparable to the shear wave travel time across a diameter. But the important thing to notice about this equation is that, with inclusion of the gravity term, density enters the equation for τ independently of the travel time or velocity term. This is the principle that allows the broad-scale density profile of the Earth to be obtained from spheroidal mode periods. For a fluid body ($\mu = 0$), the gravity term is the only one on the right-hand side of Eqs. (16.60) and (16.61). The Sun is such a body and observations of its free modes (helio-seismology) provide information about its internal structure, but in this case the assumption of homogeneity is a very poor one.

We can use a similar analysis to calculate the period of the ${}_0T_2$ mode for a homogeneous sphere. The torsional modes involve no radial motion and so give no independent information about density, but this means that they give details of the μ/ρ structure without the complication of $G\rho$ terms, such as appear in Eqs. (16.60) and (16.61), and so help in isolating the gravitational effect on the spheroidal modes. ${}_0T_2$ is the simplest torsional mode, being a twisting motion between two hemispheres, with the equatorial plane of the motion as a node. The angular displacement increases linearly with distance from this plane in the manner of the first degree zonal harmonic $\mathrm{P}_1(\cos\theta) = \cos\theta = z/R$. Dividing the sphere into discs of radii $r = R\sin\theta$ and thickness $\mathrm{d}z = R\sin\theta\,\mathrm{d}\theta$, as in the calculation of E_z above, each disc has a mass $dm = \pi R^3 \rho \sin^3\theta\, d\theta$ and moment of inertia $dI = (1/2)r^2 dm = (\pi/2)R^5\rho\sin^5\theta\, d\theta$. Taking the maximum angle of twist at the pole of the motion to be ψ, and therefore $\psi z/R = \psi\cos\theta$ at co-latitude θ, the maximum angular velocity of an elementary disc about its axis is $\omega\psi\cos\theta$ and its kinetic energy is $(1/2)dI(\omega\psi\cos\theta)^2$. Thus, the total kinetic energy is

$$\begin{aligned} E_K &= \frac{1}{2}\int_0^\pi (\omega\psi\cos\theta)^2 dI \\ &= \frac{\pi}{4}R^5\rho\psi^2\omega^2\int_0^\pi \cos^2\theta\sin^5\theta\, d\theta \\ &= \frac{4\pi}{105}R^5\rho\psi^2\omega^2. \qquad (16.62) \end{aligned}$$

Now we calculate the strain energy at an instant of maximum twist between the hemispheres. Each disc, of thickness $dz = R\sin\theta\, d\theta$, is twisted by an angle $d\psi = \psi dz/R = \psi\sin\theta\, d\theta$. We can use this in the standard formula for the torsional constant of a rod of radius r and length l, that is the torque per unit of angular twist, $\alpha = (\pi/2)r^4\mu/l$, with $l = dz = \sin\theta\, d\theta$, and write the strain energy as

$$dE_S = (1/2)\alpha(d\psi)^2 = (\pi/4)R^3\mu\psi^2\sin^5\theta\, d\theta, \qquad (16.63)$$

which integrates to

$$E_S = (4\pi/15)R^3\mu\psi^2. \qquad (16.64)$$

Now we can equate Eqs. (16.62) and (16.64) to give

$$\omega^2 = 7\mu/\rho R^2. \qquad (16.65)$$

Rewriting this as the oscillation period, τ, in terms of the travel time of a shear wave across a diameter, $T = 2R(\rho/\mu)^{1/2}$,

$$\tau = (\pi/\sqrt{7})T = 1.19T. \qquad (16.66)$$

This is very close to the ${}_0S_2$ period if the gravitational term in Eq. (16.61) is neglected. With inclusion of this term the ${}_0S_2$ mode would be faster, but for the Earth it is significantly slower, at least partly because radial motion in the fluid core imparts inertia to the system, but there is little corresponding slowing of ${}_0T_2$.

16.7 The moment tensor and synthetic seismograms

Calculation of the normal modes of the Earth is similar to calculating those of a stretched string (or rod), with the one-dimensional geometry replaced by the three-dimensional geometry of the Earth. In the case of the string, the eigenfunctions are sinusoids in space that satisfy the boundary condition of zero displacement at the ends, between which there is an integral number of half wavelengths, allowing a series of discrete frequencies. For the Earth the corresponding boundary conditions are zero tractions on the surface. The eigenfunctions or normal modes are spherical Bessel functions that describe the radial variation, and spherical harmonics for variation over spherical surfaces. Each mode has a distinct frequency determined by the boundary condition. In the general case, the displacement field can be expressed as a weighted sum of normal modes with weights adjusted to satisfy the initial conditions of excitation. For an earthquake the normal mode displacements at time zero must match the dislocation described by the earthquake moment tensor, which, as discussed in Chapter 14, can be expressed in terms of force couples. The displacement from a force couple was obtained by differentiating the displacement field from a point force with respect to the source location (Eq. (14.9)). Thus, if we solve for the normal modes generated by application of a point force within the Earth, we can synthesize the modes generated by any arbitrary moment tensor by taking appropriate derivatives and superimposing the solutions. The resulting displacement field gives synthetic seismograms at any place and time on the globe with due account for sphericity and variation with depth.

We illustrate the excitation of normal modes by a point force by examining the one-dimensional case of longitudinal oscillations of an elastic rod, fixed at $x = 0$ and $x = L$, with displacement u in the x-direction (Fig. 16.12). The normal modes are given by

$$u_n = \sin(\omega_n x/c)\exp(-i\omega_n t), \qquad (16.67)$$

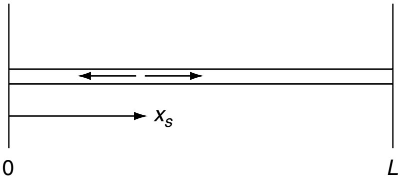


FIGURE 16.12 Geometry of a force doublet excitation of a clamped rod.

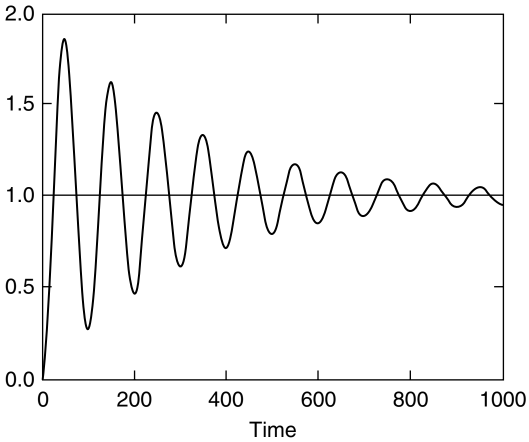


FIGURE 16.13 Time variation of a normal mode excited by a source that is a step function at $t = 0$.

with eigenfrequencies

$$\omega_n = n\frac{\pi c}{L}, n = 1, 2, 3, \ldots, \tag{16.68}$$

where $c = \sqrt{E/\rho}$ is the velocity that satisfies the one-dimensional wave equation, assuming that the wavelength is much greater than the rod thickness,

$$E\partial^2 u/\partial x^2 = \rho\partial^2 u/\partial t^2, \tag{16.69}$$

and E and ρ are Young's modulus and density. Then

$$c^2\partial^2 u/\partial x^2 = \partial^2 u/\partial t^2. \tag{16.70}$$

The effect at time $t \geq 0$ of a disturbance at time $t = 0$, expressed as a weighted sum of normal modes, is

$$u(x,t) = \sum_i a_n u_n(x,t) = \sum_i a_n \sin(k_n x) \times (1 - \exp(-\mathrm{i}\omega_n t)), \tag{16.71}$$

where $k_n = \omega_n/c$ and the mode shapes are orthogonal, so that

$$\frac{2}{L}\int_0^L \sin\left(\frac{n\pi x}{L}\right)\sin\left(\frac{m\pi x}{L}\right)\mathrm{d}x = \delta_{mn} = 1 \quad \text{if } m = n, \text{ or} = 0 \text{ if } m \neq n, \tag{16.72}$$

where m and n are integers. Suppose that at time $t = 0$ and location $x = x_s$ a force f_0 is applied in the x-direction and is maintained indefinitely. It can be represented by

$$f = f_0\frac{\delta(x - x_s)}{a}H(t), \tag{16.73}$$

where f is the force per unit volume, a is the cross-sectional area of the rod, $\delta(x)$ is a delta function in x, and $H(t)$ is a Heaviside step function. The equation of motion (Eq. (16.69)) becomes

$$E\partial^2 u/\partial x^2 + f_0\frac{\delta(x - x_s)}{a}H(t) = \rho\partial^2 u/\partial t^2. \tag{16.74}$$

Substitution of Eq. (16.71) for u and solving for the coefficients a_n in that equation by taking a Fourier transform and using Eq. (16.72) to resolve individual coefficients (Problem 16.3) gives

$$u(x,t) = \sum_i \frac{2f_0}{M}\sin(k_i x_s)\sin(k_i x)\frac{1 - \cos\omega_i t}{\omega_i^2}, \tag{16.75}$$

where $M = aL\rho$ is the mass of the rod. Equation (16.75) can be interpreted as displacement equals a sum of modes, evaluated at the point of application of the force, multiplied by the eigenfunctions and the time variation (Fig. 16.13).

We showed in Chapter 14 that dislocations are equivalent to force couples and doublets. In the rod case a force doublet corresponds to clamping a small section of the rod at x_s in compression and releasing it at $t = 0$. The solution for a force doublet is obtained by placing a second oppositely directed force a distance dx from x_s and summing the solutions. For this case the solution is obtained by adding to Eq. (16.75) the displacement caused by a second, oppositely directed force at distance dx from x_s. In effect this means differentiating Eq. (16.75) with respect to x_s and multiplying by dx, and is completely equivalent to the differentiation of Eq. (14.8) to produce Eq. (14.9). Following the procedure in Section 14.2, we replace the force moment $f_0 dx$ with dislocation moment $Eab = M_0$, where b is the dislocation and a the area, to obtain

$$u(x,t) = \frac{M_0}{M} \sum_i 2k_i \cos(k_i x_s) \sin(k_i x) \frac{1 - \cos \omega_i t}{\omega_i^2}, \tag{16.76}$$

where, as in Eq. (14.9), we have the moment term, M_0, multiplied by the Green's function, here expressed as a sum of normal modes. If we include attenuation,

$$u(x,t) = \frac{2M_0}{M} \sum_i k_i \cos(k_i x_s) \sin(k_i x) \times \frac{1 - \exp[-\omega_i t/2Q_i] \cos \omega_i t}{\omega_i^2}. \tag{16.77}$$

Note that $k_i \cos(k_i x_s)$ is the strain $\partial u/\partial x$ of the ith mode or $e^i_{xx}(x_s)$. Then

$$u(x,t) = \frac{2M_0}{M} \sum_i e^i_{xx}(x_s) \sin(k_i x) \times \frac{1 - \exp[-\omega_i t/2Q_i] \cos \omega_i t}{\omega_i^2}. \tag{16.78}$$

Thus, the weights are equal to the spatial derivatives of the eigenfunctions, that is, strains evaluated at the source point x_s, multiplied by the corresponding elements of the moment tensor and divided by mass. Equation (16.78) describes a force doublet applied to a rod. To use this approach to describe a dislocation in the Earth, the one-dimensional eigenfunctions are replaced with three-dimensional ones. So, based on Eq. (16.78), we can write down the solution directly. The displacement field $u(x,t)$ from a point source earthquake with arbitrary moment tensor $M_{pq}(x_s)$ (Aki and Richards, 2002), represented as a sum of normal modes, is

$$u(x,t) = \sum_i \left[e^i_{pq}(x_s) M_{pq}(t) \right] u^i(x) \times \left(\frac{1 - \exp(-\omega_i t/2Q_i) \cos \omega_i t}{\omega_i^2} \right), \tag{16.79}$$

where $u^i(x)$ is the displacement attributed to the ith normal mode and $e^i_{pq}(x_s)$ is the corresponding strain evaluated at the location of the source. Once the normal modes are calculated, Eq. (16.79) can be used to calculate the seismic wave field for an arbitrary moment tensor, M_{pq}. The term on the RHS of Eq. (16.79) represents a damped sinusoidal oscillation with offset $1/\omega_i^2$ (Fig. 16.13) and so includes the static displacement (offset) of the dislocation. The normal modes, $u^i(x)$, depend on the Earth model, and are found by satisfying the equation of motion and the boundary conditions of continuity of stress and displacement at each layer boundary including the surface. Calculation of the modes has progressively improved with refinements to take into account gravity, the rotation of the Earth, its ellipticity, and lateral heterogeneity.

Equation (16.79) is used in inversion of seismic data to determine M_{pq}. This is the basis of the Harvard moment tensors that have been calculated worldwide since about 1977 for earthquakes above magnitude 5.6. By adding a sufficient number of normal modes one obtains all the transient waves that radiate from a source, including all body waves and surface waves (Fig. 16.14). In that they can all be represented as normal modes, the distinction between body and surface waves and free oscillations is a matter of timing as to when the observations are made. The transient waves can only be observed early in the record, while they remain isolated from other phases. As time progresses multiple reflections and transmissions around the globe and dispersion of wave trains cause the energy to merge into a continuous global vibration, which is better analysed in the frequency domain for mode peaks and attenuation. A fast method of calculating modes (Woodhouse 1983, 1988) is capable of modelling body wave phases at frequencies up to 0.16 Hz. However, in practice the moment tensor obtained from normal modes is largely made up of waves with periods much longer than the earthquake duration, and so is representative of the moment distribution averaged in time and space, referred to as the centroid moment tensor. In contrast, the moment determined from first-arriving P-body waves measures fault motion at the hypocentre, which, unless the event expands symmetrically, will be offset from the centroid.

Normal mode eigenfunctions (Fig. 16.15) show how the different frequencies give a weighted

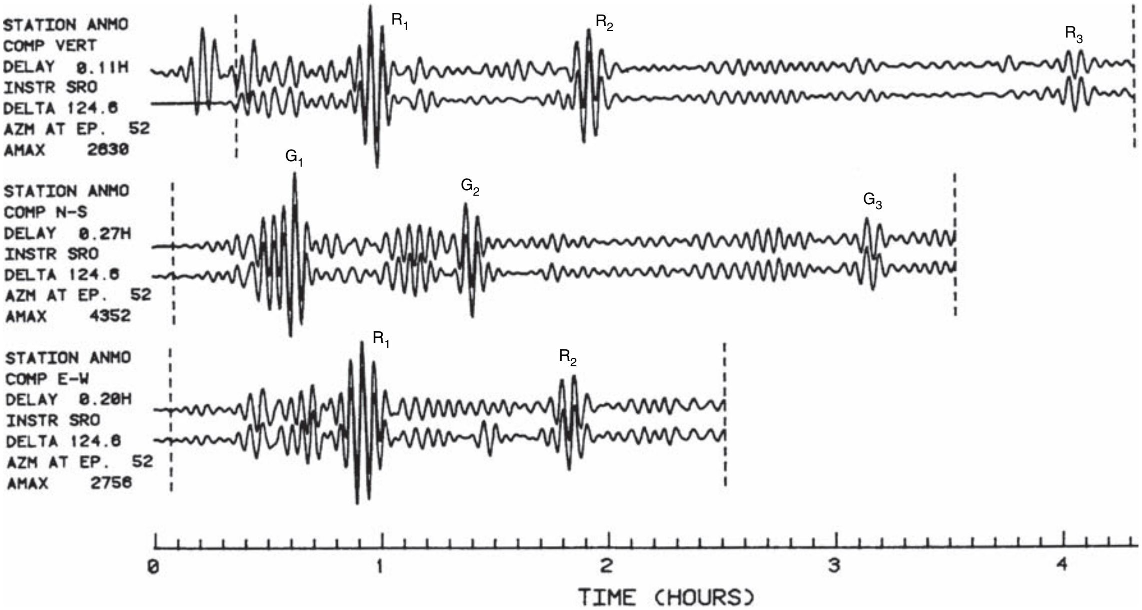


FIGURE 16.14 Synthetic seismograms from normal mode summations compared with observations. The top traces are the observations and the bottom traces the synthetics for three earthquakes measured at the seismic station ANMO in Albuquerque, New Mexico. From Woodhouse and Dziewonski (1984). R_1, R_2, . . . are Rayleigh waves from the short and great circle paths from the source to the station. G_1, G_2, . . . are the corresponding Love waves.

sampling of the Earth's internal structure. From Eq. (16.75) we see that the excitation of a mode depends on its strain calculated at the location of the earthquake. Earthquakes near the surface excite fundamental modes which have their eigenfunctions concentrated near the surface. Deep earthquakes excite higher-order modes or overtones with weaker excitation of fundamental modes.

Of particular interest are the modes that have significant energy in the inner core, and have played an important role in confirming the solidity of the inner core. As early as 1946 Bullen used the amplitudes of reflected P-waves from the inner core surface, but the inference that it represents a liquid–solid interface required several assumptions. The reflection coefficient depends on the impedance ($V_P\rho$) contrast (Eq. (16.29)). Independent evidence from travel times for P-waves suggested that the contrast was caused by a significant increase in V_P from outer to inner core rather than a density change. The jump in V_P could be explained if the inner core is solid, so that the P-wave modulus changes from $\chi = K_{\text{outer}}$ to $\chi = K_{\text{inner}} + \frac{4}{3}\mu$ (Eq. (16.1)). Alternatively, if the inner core remained liquid the required change in K was unreasonably large, and a decrease in density with depth is implausible. However, the definitive test required identification of inner core shear waves. Because the outer core is fluid, the shear waves have to be generated by conversions of P-waves at the inner core boundary. Such conversions are inefficient because incident angles are small. Furthermore, as we mention below, S-waves appear to be highly attenuated in the inner core. After passage though the inner core they convert back to P-waves, which must be distinguished from other phases that might arise from heterogeneities in the mantle. Thus, early reports of inner-core S-body wave phases (such as PKJKP) have been questioned to the point that skepticism was expressed as to whether they would ever be observed in high frequency body waves (Doornbos, 1974). Analysis of normal modes can overcome these problems, because, rather than relying on identification of individual weak S phases, the mode energy is

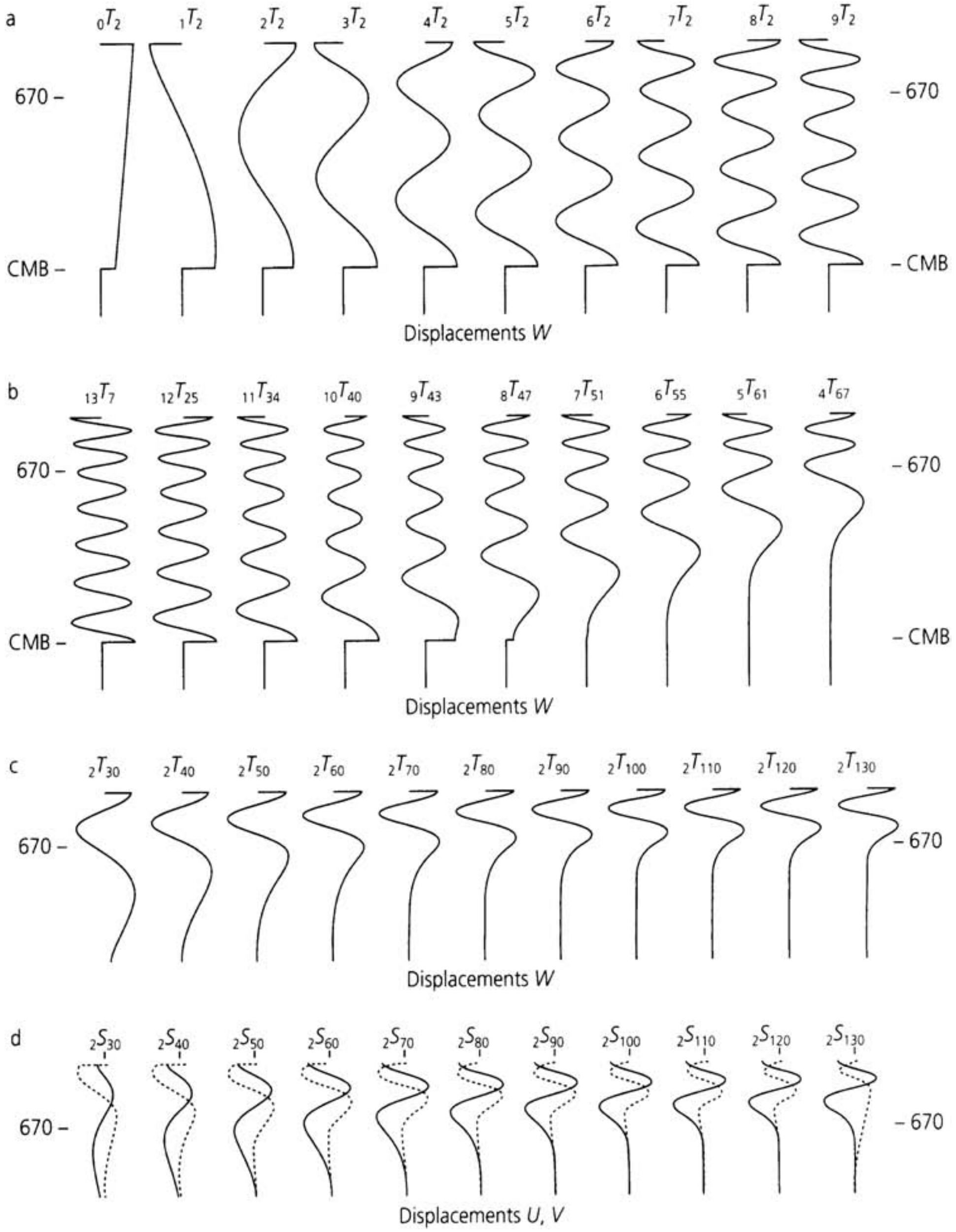


FIGURE 16.15 Eigenfunctions of toroidal and spheroidal normal modes from the surface to the core–mantle boundary (Dahlen and Tromp, 1998). (a), (b) and (c) show torsional modes. (d) shows spheroidal modes, with solid lines representing radial displacements and dashed lines tangential displacements.

effectively stacked by spectral analysis of long wave-trains. Modes having eigenvectors with significant particle motions in the inner core ($_6S_2$ and $_7S_3$) indicated an average $V_S = 3.52 \pm 0.03\,\mathrm{km/s}$ (Masters and Gilbert, 1981). This apparently low value is expected for solid iron at inner-core pressure, as explained in Section 18.8. Now that anisotropy is recognized (Section 17.7), we know that the inner core is crystalline and aligned.

The usefulness of mode-generated synthetic seismograms was illustrated in a resolution by Deuss *et al.* (2000) of the debate over identification of phases propagating as shear waves in the inner core (in the notation illustrated by Fig.17.7, PKJKP, SKJKP and also pPKJKP, which has a surface reflection from a deep event). These phases have very small amplitudes because the necessary P to S and S to P conversions of the inner core boundary are weak; additionally Doornbos (1974) pointed out that the low *Q* for S-waves would make them completely unobservable at high frequencies ($\sim$1 Hz) and reported observations have been disputed. Using a method of phase stacking records from the deep 1996 Flores Sea event, Deuss *et al.* (2000) compared observed arrivals at the expected times with synthetic seismograms for the PREM model. They repeated the calculation of synthetics for a modified PREM, with no inner core shear wave but the same V_P. By accepting only arrivals seen with the solid inner core model but not the fluid model, they were able to discriminate against interfering minor signals from the mantle. Superimposed pPKJKP and SKJKP arrivals survived this test, but a similar analysis for a deep 1994 Bolivia event found no unambiguous evidence for inner-core S-waves. Although the Flores Sea event provided the only body wave evidence for inner-core rigidity, it confirmed the average $V_S \approx 3.6\,\mathrm{km/s}$ inferred from free oscillation modelling.

17

Seismological determination of Earth structure

17.1 Preamble

Our knowledge of the Earth's internal structure would almost certainly be very primitive if there were no earthquakes. Explosive sources of elastic waves would have been recognized and it is possible that exploration seismology, directed to the identification of oil-bearing structures in the crust, would have been well developed, but it is likely that deep-Earth seismology would have been rudimentary. The larger nuclear weapons tests generate waves of sufficient amplitudes to be detected at remote stations, but recognition of the test-monitoring possibilities of seismology depended on the fact that the subject was already well developed when nuclear testing began. Even if the weapons-testing agencies had appreciated the seismic detection possibilities, a shroud of secrecy would have ensured that evidence of deep Earth structure emerged only very slowly.

This hypothetical situation emphasizes how completely our detailed knowledge of the deeper parts of the Earth relies on observations of seismic waves. Quantitative instrumental data on teleseismic waves (from distant earthquakes) were first obtained in the late nineteenth century and, as the theory of elastic waves was then already well established, seismology developed rapidly. The use of earthquake-generated waves to study the Earth's interior is a mature and sophisticated science, to which this chapter is necessarily only a brief introduction. Comprehensive treatments are by Bullen and Bolt (1985), Aki and Richards (2002) and Stein and Wysession (2003).

As we now know, the Earth's internal layers, as well as the surface, are close to oblate ellipsoids, symmetrical about the rotational axis. It is not an entirely trivial matter to establish that this is true. Major earthquakes occur in limited bands around the Earth and these bands obviously differ in some respects from the larger non-seismic areas. Also seismometers are almost all placed on land, so that there is a danger of a systematic bias in the observations. It would be very difficult to ensure that a seismological model based entirely on body waves is not affected by such bias. However, surface waves, propagating around the Earth on paths that may be predominantly either continental or oceanic, give information about lateral heterogeneity, complementary to that obtained from the body waves that penetrate the deep interior. At the lowest frequencies, free oscillations of the entire Earth are observed and they directly indicate average properties. Average models are normally represented by radial variations in properties for a sphere, although derived by assuming elliptical internal surfaces. Thus, we are confident of the reliability of recent global average Earth models, the most widely used of which is PREM (Preliminary Reference Earth Model (Dziewonski and Anderson, 1981), see Appendix F). Departures of the real Earth average from PREM are probably less significant than the local and regional variations that have now become subjects of study.

Early Earth modelling, notably the pioneering work of K.E. Bullen in the 1930s, used body wave travel times to obtain the variations of P- and S-wave speeds through the Earth (Section 16.6). These give the ratios of elastic moduli to density, and additional information was required to estimate density independently. Using the known mass and moment of inertia of the Earth with the assumption of simple adiabatic compression with depth where the wave speeds indicated that this was reasonable, Bullen obtained Earth models that are remarkably close to our present understanding. The most important additional data available for the development of recent Earth models are the periods of free oscillation (Section 16.6). The spheroidal modes, which involve radial motion, have gravitational as well as elastic restoring forces and so give independent evidence of density structure. There are also torsional or toroidal modes that give strong control on the shear wave structure in Earth modelling.

To account for the properties of the deep regions of the Earth, we must allow for the changes in these properties that are caused by the high pressures. For a homogeneous layer, seismology itself provides a method of doing this. The ratio $K_S/\rho = (\partial P/\partial \rho)_S = (V_P{}^2 - (4/3)V_S{}^2)$ is obtained directly from the wave speeds and if density everywhere is adequately modelled, so that gravity variation is determined by the model, then so is the density gradient (Section 17.5). This is most precisely true for the outer core, for which homogeneity is assured by three-dimensional stirring at speeds of tens of kilometres per year, as evidenced by the geomagnetic secular variation (Section 24.3). This ensures not just homogeneity but an adiabatic temperature gradient, so that compression is described by the adiabatic modulus, K_S, derived from the wave speed. Theories that account for the strong variations in density and bulk modulus with pressure are a subject of Chapter 18.

The broad-scale layering of the Earth, outlined by seismology, represents the average and stable state of the internal structure. Finer details, lateral variations and anisotropy, are attributed to the dynamic behaviour and must change slowly with time. They are most clearly observed in the crust and uppermost mantle but are recognized to occur at all depths in the mantle. This is a new frontier of the subject. It is obvious that the Earth must be laterally heterogeneous in the uppermost 100 km or so where surface waves propagate, because surface waves from an earthquake are not sharply re-focussed to cause damage at an anti-focus on the opposite side of the Earth. But heterogeneities occur throughout the mantle and are generally believed to be related to the tectonic pattern. Observations become less detailed with depth, and in the lower mantle we are confident that there are features such as plumes and fragments of subducted slabs that have not yet been resolved seismologically. However, in the D″ layer, the lowermost 200 km or so of the mantle, strong lateral heterogeneities are well documented. In Chapter 12 these are referred to as crypto-continents and crypto-oceans, by analogy with the surface structure (see Fig. 12.3), but are probably to be explained, at least partly, by a phase transition to the post-perovskite mineral structure, referred to in Section 2.7.

The ellipticity of the core is not well constrained by seismological observations and this is even more true of the inner core. Evidence from very long baseline interferometry (VLBI) observations of the nutations suggest that the core is more elliptical than equilibrium theory suggests, as mentioned in Section 7.5. There is a strong reason to expect a non-equilibrium ellipticity of the inner core. It is likely that solidifying material is preferentially deposited on the equator, giving the inner core an excess ellipticity from which it deforms towards its equilibrium flattening. Crystal alignment resulting from the deformation is the most plausible explanation for the inner core anisotropy (Yoshida *et al.*, 1996; see Section 17.9).

17.2 Refraction in a plane layered Earth

For seismic wave propagation over distances that are small compared with the radius of

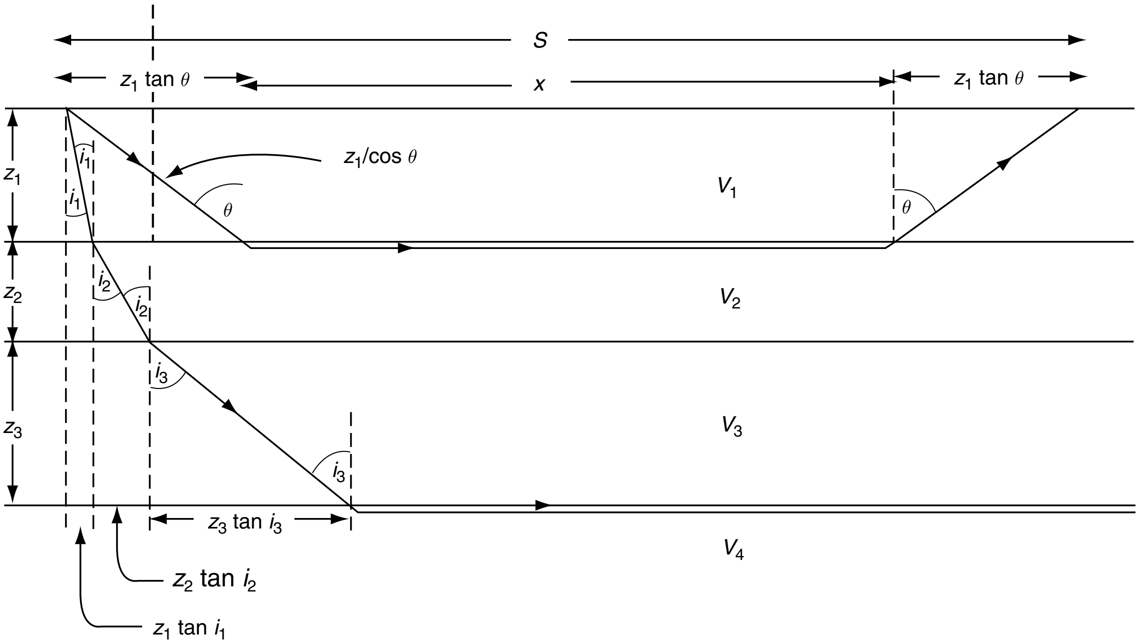


FIGURE 17.1 Geometry of seismic rays in a plane layered Earth model, having P-wave speed, V_P, progressively increasing with depth. Waves that travel for much of their paths in deeper and faster layers, as illustrated, are known as head waves.

the Earth, the sphericity has little effect on travel times. A model in which wave speed increases with depth in a series of plane, horizontal layers is a useful approximation to reality in some situations. It is also a convenient starting point for a more general discussion of travel times.

Consider the model in Fig. 17.1, in which two of the possible seismic rays from an earthquake or explosion are shown. The shallower ray just penetrates layer 2, being refracted horizontally within it. Since this means that the angle of refraction from the boundary with layer 1 is 90°, Snell's law (Eq. (16.5)) gives the angle of incidence, θ, in terms of the ratio of wave speeds in the two layers,

$$\sin\theta = V_1/V_2. \tag{17.1}$$

Then the distance travelled in layer 2, at speed V_2, is related to the total distance, S, and the thickness, z_1, of layer 1,

$$x = S - 2z_1 \tan\theta. \tag{17.2}$$

The total travel time for this ray is the sum of times in the two layers

$$T = \frac{x}{V_2} + \frac{2z_1}{V_1 \cos\theta}, \tag{17.3}$$

so that, substituting for x by Eq. (17.2), we obtain the $T(S)$ relationship for a family of rays that penetrate layer 2 but no deeper and travel to various distances S:

$$T = \frac{S}{V_2} + 2z_1\left(\frac{1}{V_1\cos\theta} - \frac{\tan\theta}{V_2}\right). \tag{17.4}$$

With substitution for θ by Eq. (17.1), this becomes

$$T = \frac{S}{V_2} + 2z_1 \cdot \frac{\left(V_2^2 - V_1^2\right)^{1/2}}{V_1 V_2}. \tag{17.5}$$

For this family of rays the travel time is linear in S, at speed V_2, but with an intercept due to layer 1.

The analysis is readily extended to a multiplicity of layers. If the deepest layer penetrated by a family of rays has speed V_n, then this is the speed across the surface of the wave arrivals of this family, and the angles to vertical of the rays in each of the higher layers are

$$\sin i_1 = V_1/V_n; \;\; \sin i_2 = V_2/V_n, \ldots \tag{17.6}$$

Using these angles to write down the travel time and horizontal distance of travel for each of the layers, as for the single layer, we arrive at the general result

$$T = \frac{S}{V_n} + 2\sum_{j=1}^{n-1} \frac{z_j\left(V_n^2 - V_j^2\right)^{1/2}}{V_n V_j}. \tag{17.7}$$

As before, this is a linear relationship, but with an intercept that depends on the depths and speeds of all of the higher layers. Thus, to determine these parameters from a listing of travel times, as in Problem 17.2 of Appendix J, one can first obtain all the speeds from the series of gradients of the travel-time curve, and then use the intercepts to find the layer thicknesses in turn, starting from the top.

In such an exercise it is necessary to note that travel times from a near source generally give only the time of travel of the first arriving pulse at each of a series of recording stations. Thus, for the nearest stations the first wave has travelled in the top layer and only beyond a certain distance does a ray that has reached layer 2 arrive first. It must travel a sufficient distance in the faster layer to compensate for the longer path (see Problem 17.3). There are corresponding breaks in the travel-time curve with a gradient change at each point where the faster path extends to a new layer (Fig. 17.2).

The horizontally layered model is sufficiently close to the real Earth on a regional scale that a single line of refraction data can give satisfactory values of layer thicknesses and speeds. If layers are inclined, then not only are depths ambiguous but wave speeds are biased. Consider a layer of speed V_1 overlying one with speed V_2, and having an inclined boundary between them, dipping downwards at angle θ away from the source of seismic waves. The slowness across the surface, that is, the inverse of surface speed, of a head wave that travels in the lower medium is (Problem 17.4)

$$\frac{dT}{dS} = \frac{\cos\theta}{V_2} + \frac{\sqrt{V_2^2 - V_1^2}}{V_1 V_2}\sin\theta. \tag{17.8}$$

For small θ, the first term is little different from $1/V_2$, as observed with horizontal layers. The second term is normally the principal cause of a biased result. The problem is overcome by making measurements of propagation in the opposite direction across the same area. Then the $\sin\theta$ term is reversed in sign. Both V_2 and θ can be found with a reversed profile, as when explosive sources are used in exploration, but the method has no application to earthquake studies. dS/dT, from Eq. (17.8), is referred to as the apparent velocity of the refracted wave along the surface. If the layer is upward sloping (θ negative), for the case $\sin\theta = V_1/V_2$, the apparent velocity can be infinite, with all refracted pulses arrive at the surface at the same time.

For exploration of the upper crust, much greater detail is required. This is usually obtained by using reflections rather than refractions. For reflection studies sources and receivers are generally closer together than in refraction work, so that reflections are obtained close to normal incidence. Two-way travel times are measured and this translates into boundary depths when the layer speeds are known. The travel time of a reflection from a horizontal layer at depth h as a function of distance, S, is

$$T = \frac{1}{V_1}\sqrt{4h^2 + S^2}, \tag{17.9}$$

which is the equation for a hyperbola. This distinguishes reflected arrivals from refracted ones on a travel time (T–S) plot since refracted arrivals form straight lines (Eq. (17.7)).

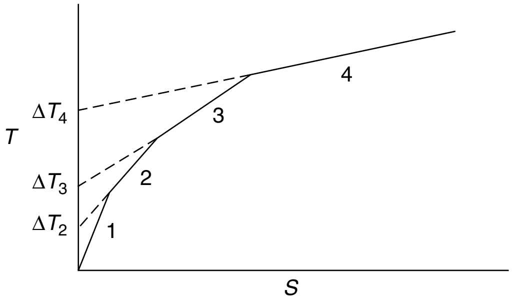


FIGURE 17.2 Travel-time curve for first arriving pulses from a layered structure, as in Fig. 17.1. Numbers on segments indicate the deepest layer penetrated by each 'family' of rays.

17.3 Refraction in a spherically layered Earth

Consider a spherical model of the Earth in which wave speed, V, varies radially. For the purpose of the argument that follows it is convenient to consider a layered model, with each layer uniform, but this restriction can be removed by introducing an infinite number of infinitesimal layers with a graded wave speed. The geometry of a seismic ray through a three-layer model is shown in Fig. 17.3. Applying Snell's law (Eq. (16.5)) to each of the boundaries, A and B,

$$\left.\begin{aligned}\frac{\sin i_1}{V_1} &= \frac{\sin f_1}{V_2}\\ \frac{\sin i_2}{V_2} &= \frac{\sin f_2}{V_3}\end{aligned}\right\}. \tag{17.10}$$

From the two triangles,

$$q = r_1 \sin f_1 = r_2 \sin i_2, \tag{17.11}$$

so that if we multiply Eq. (17.9) by r_1 and Eq. (17.10) by r_2 we can then equate them to one another,

$$\frac{r_1 \sin i_1}{V_1} = \frac{r_1 \sin f_1}{V_2} = \frac{r_2 \sin i_2}{V_2} = \frac{r_2 \sin f_2}{V_3}. \tag{17.12}$$

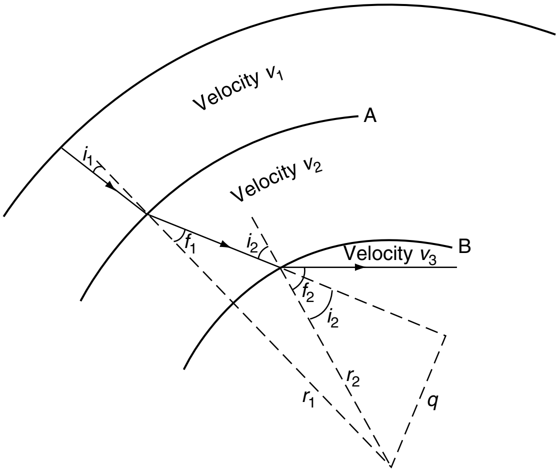


FIGURE 17.3 Path of a seismic ray through three uniform layers of a spherical model, showing the geometrical construction used to prove that the seismic ray parameter, $p = r \sin i/v$, remains constant along the path.

Equation (17.12) could be extended to any number of boundaries or to gradual refraction in a layer of progressively increasing speed. It demonstrates that

$$\frac{r \sin i}{V} = p \tag{17.13}$$

is a constant for the ray, where i is the angle between the ray and the radius at any point. p is referred to as the ray parameter. It remains constant not only at refractions and reflections, but also at wave conversions (P to SV or vice versa). By determining the parameter of a ray we obtain the value of r/V at its point of deepest penetration, where $i = 90°$.

Now consider two rays from a common surface source with infinitesimally different ray parameters, as in Fig. 17.4. Their distances of travel are represented by the angles Δ and $(\Delta + \mathrm{d}\Delta)$, subtended at the centre of the Earth by the ends of the ray paths. PN is the normal from P to SQ and is therefore a wavefront so that the travel times of paths SP and SN are equal. The path difference between the rays is therefore QN and the travel time difference is

$$\mathrm{d}T = \mathrm{QN}/V_0, \tag{17.14}$$

where V_0 is the wave speed in the surface layer. But

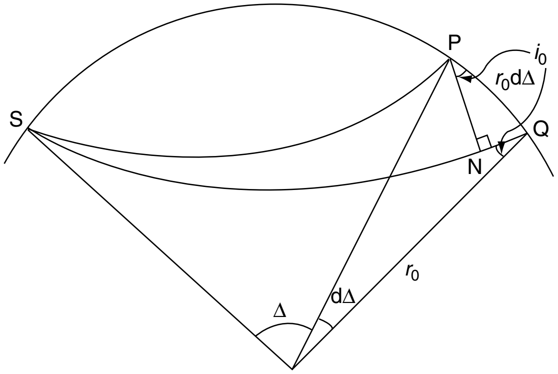


FIGURE 17.4 Two seismic rays with infinitesimally different distances of travel. This geometrical construction is used to relate the ray parameter, p, to the slowness (1/speed) with which wave arrivals cross the surface. p is the gradient of the travel time curve.

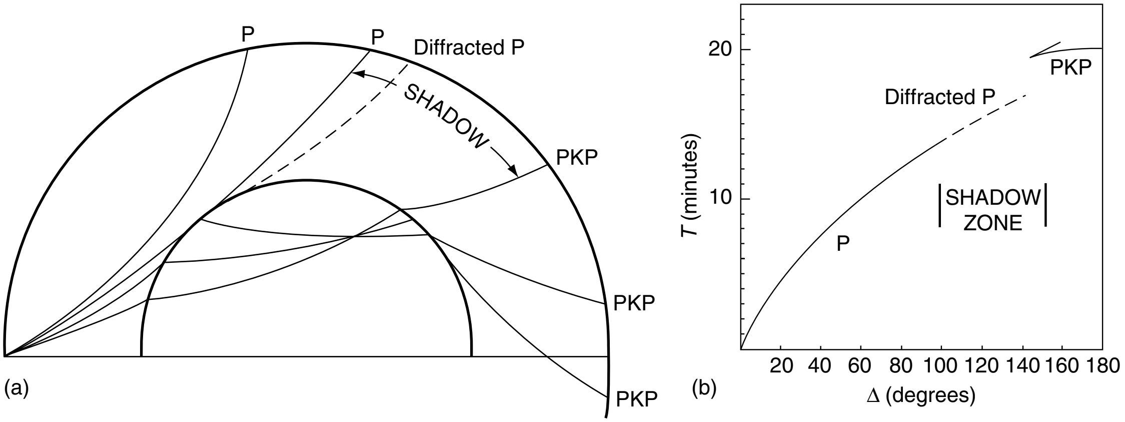


FIGURE 17.5 (a) Seismic rays refracted by the core, causing a shadow zone for direct P waves. (b) The corresponding travel time curve has a break. PKP is the nomenclature of a P wave having part of its path in the core.

$$\mathrm{QN} = \mathrm{PQ}\sin i_0 = r_0 \mathrm{d}\Delta \sin i_0, \tag{17.15}$$

so that

$$\frac{\mathrm{d}T}{\mathrm{d}\Delta} = \frac{r_0 \sin i_0}{V_0} = p. \tag{17.16}$$

Thus, a ray parameter is the gradient of the travel time curve, or slowness, and so can be directly observed as a function of total angular distance of travel, Δ. Linked arrays of seismometers give precise observations of $\mathrm{d}T/\mathrm{d}\Delta$ and are particularly useful in investigating complications, such as occur in the mantle transition zone.

If wave speed increased gradually inwards throughout the Earth, then p would progressively decrease with Δ, and $p(\Delta)$ would be a continuous, monotonic function. There are large depth ranges over which this regular behaviour occurs, giving correspondingly straightforward p–Δ and therefore T–Δ relationships. However, two effects cause breaks. Whereas an increase in speed with depth causes rays to turn upwards, a decrease causes them to turn downwards and, if this is more than a slight effect (corresponding to the curvature of a level surface at that depth), there is a range of depths over which no rays have their points of deepest penetration and a break appears in the travel time curve. The necessary condition for this to occur may be obtained from Eq. (17.13). Since p is constant,

$$p\, \mathrm{d}V/\mathrm{d}r = \sin i + r\cos i\, \mathrm{d}i/\mathrm{d}r. \tag{17.17}$$

Substituting for sin i from Eq. (17.13) and rearranging,

$$\frac{\mathrm{d}i}{\mathrm{d}r} = \frac{p}{r\cos i}\left(\frac{\mathrm{d}V}{\mathrm{d}r} - \frac{V}{r}\right). \tag{17.18}$$

Thus, the condition to be satisfied to avoid a decrease in i with depth and a break in the travel-time curve is $\mathrm{d}V/dr < V/r$. At the core–mantle boundary the P-wave speed drops to less than 60% of its value at the base of the mantle, causing sharp downward refraction of P-waves and leading to a shadow zone of distances over which direct P-waves are not observed (Fig. 17.5).

A sharp increase in wave speed with depth causes a complication of a different kind. The normal trend is an increase in Δ with increasing dip angle of the rays, that is, decreasing angle of incidence, i_0. Since $p \propto \sin i_0$, this means that Δ increases with decreasing p. But the trend is reversed by sharp refraction where the rays enter a steep velocity gradient; the normal trend is restored with further decrease in p, that is, increasing steepness of the rays. Such a velocity increase, although a complicated one, occurs in the transition zone in the mantle, causing a range of distances over which three direct P wave arrivals may be observed (Fig. 17.6). This is known as triplication.

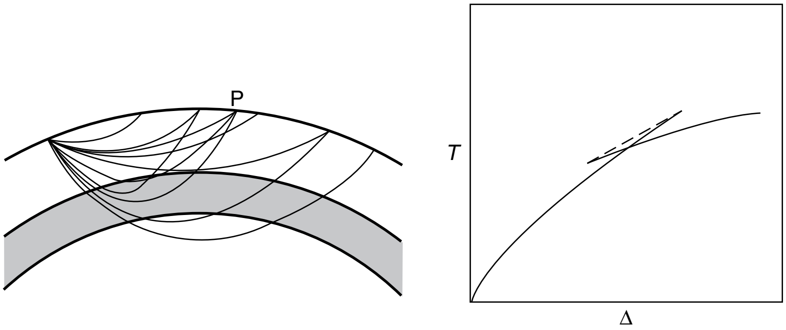


FIGURE 17.6 A zone in which velocity increases sharply with depth causes reversal of the p–Δ trend and a range of distances over which arrivals are triplicated. A point such as P may receive three rays, each one on a different limb of the $T-\Delta$ curve. This figure is illustrative and not to scale.

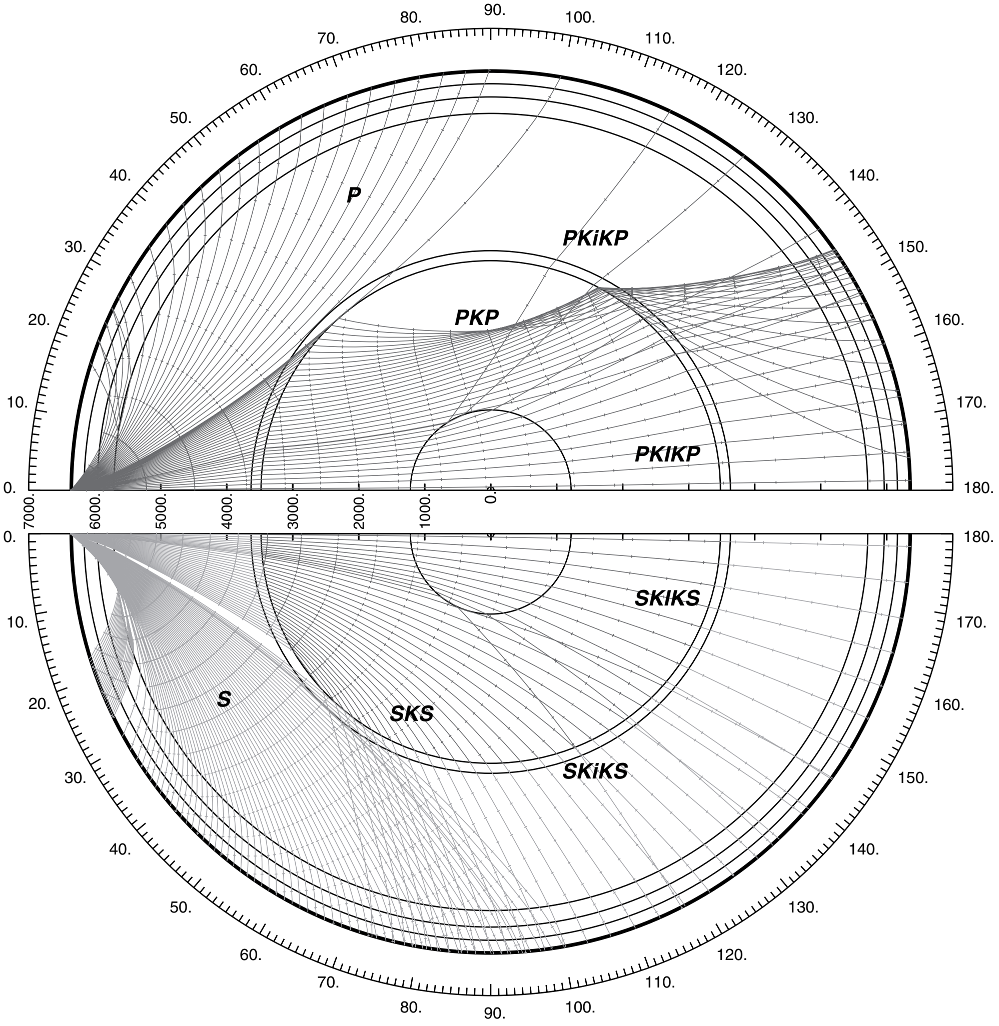


FIGURE 17.7 Paths of seismic rays through the Earth, illustrating their nomenclature. Figure by courtesy of B. L. N. Kennett.

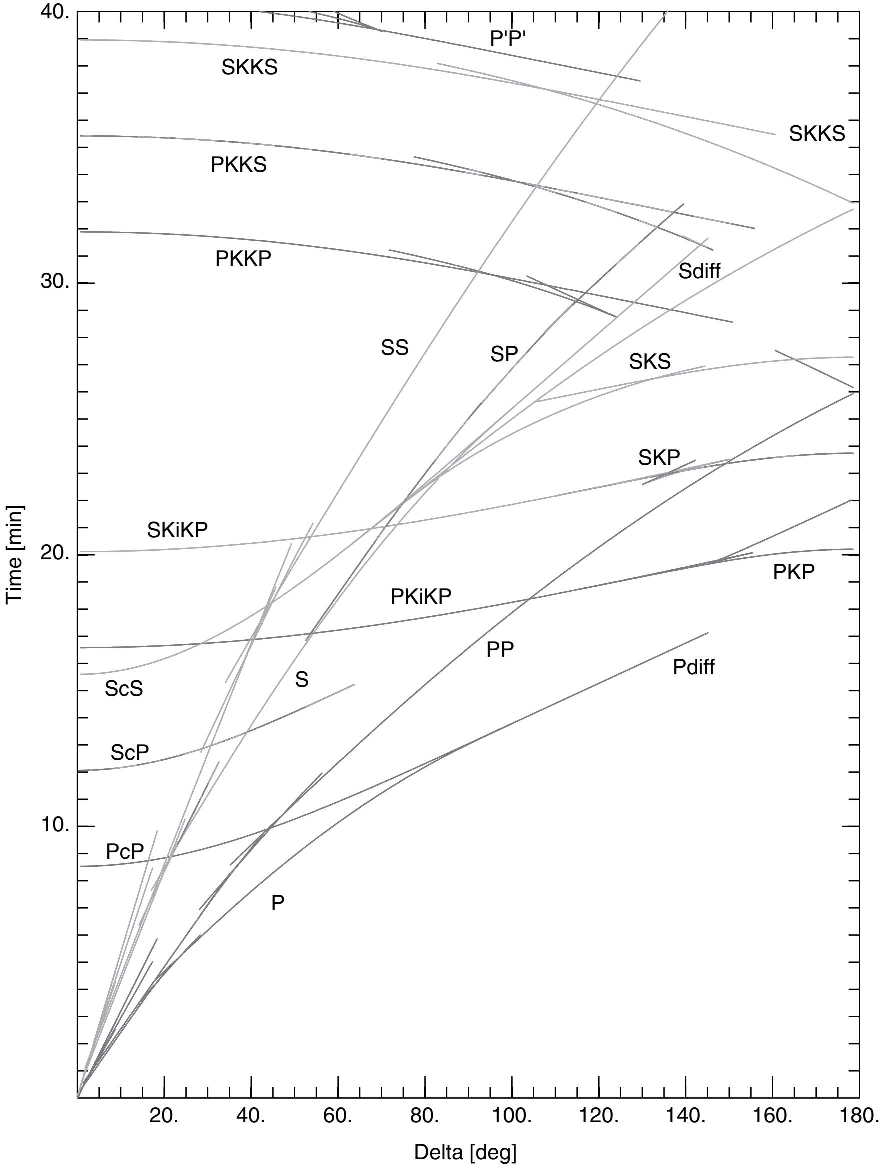


FIGURE 17.8 (a) Graph of travel times of seismic phases identified in the IASPEI 1991 Seismological Tables. Reproduced by permission from Kennett and Engdahl (1991). (b) The lower part of Fig. 17.8(a) with travel times from the Bulletin of the International Seismological Centre (ISC). Courtesy of B. L. N. Kennett.

17.4 Travel times and the velocity distribution

Complications in the structure of the Earth mentioned above cause a multiplicity of wave paths, many of which have well observed travel times and are useful in determining details of the structure. Fig. 17.7 identifies some of these, with letters P and S referring to waves in the mantle. For sources below the surface, waves are also received from surface reflections and in these cases an initial p or s is applied to the designation, e.g., pP, sS. K indicates a P-wave in the outer

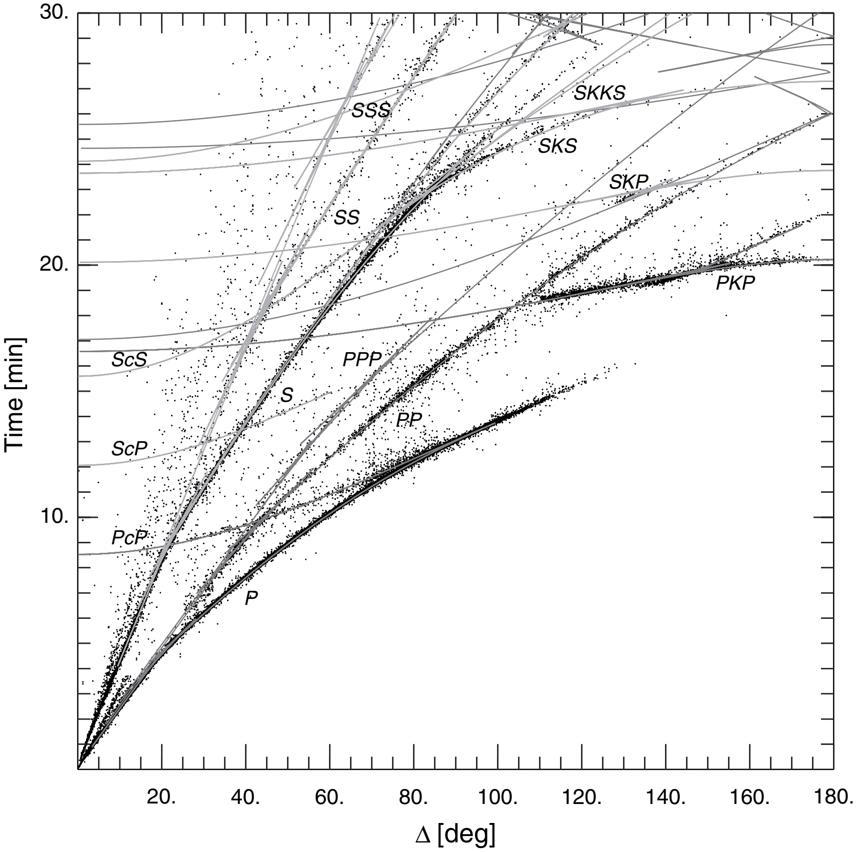


FIGURE 17.8 (Cont.)

core (German Kern or kernel), where there are no S-waves, and I refers to a P-wave in the solid inner core. The letter J is reserved for shear waves in the inner core. Although shear waves would be expected for a solid, observations are marginal (Section 16.6). Excitation of J phases requires conversion from P- to S-waves on entry to the inner core and conversion back to P-waves on re-emergence into the outer core, and these conversions are very weak. As at the surface, lower case letters indicate reflections. Occasionally numbers appear in the nomenclature. These give either the depth in kilometres at which reflections occur, or the multiplicity of internal reflections, especially at the core–mantle boundary.

For many years the tables of travel times published in 1940 by H. Jeffreys and K. E. Bullen, and known as the J-B tables, were the reference standard. Improvements in certain details have appeared from time to time, particularly direct P- and S-wave times, but there is now an updated comprehensive set of tables (Kennett *et al.*, 1995; wwwrses.anu.edu.au/seismology/ak135), illustrated in Fig. 17.8. This model, ak135, is based on an earlier model (iasp91) (Kennett and Engdahl, 1991; Montagner and Kennett 1996). The tables give times for numerous phases, as a mutually consistent set, obtained by adjusting a velocity model (Fig. 17.9) to give a best fit to a large number of observed travel times. Further important developments in modelling the velocity structure have begun to appear with the use of records from a global digital seismograph network. The records can be 'stacked', that is, added together numerically, to enhance signals, with cancellation of noise that is not coherent on the multiple records that are added. Shearer (1990)

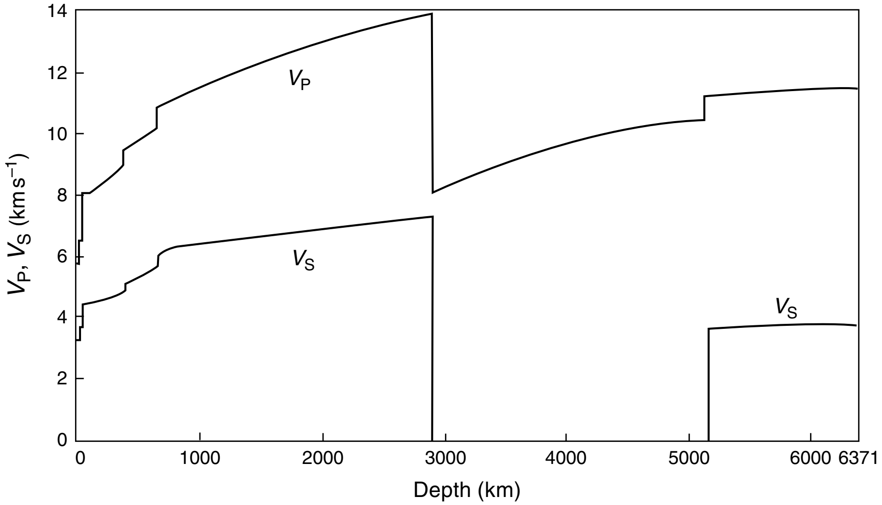
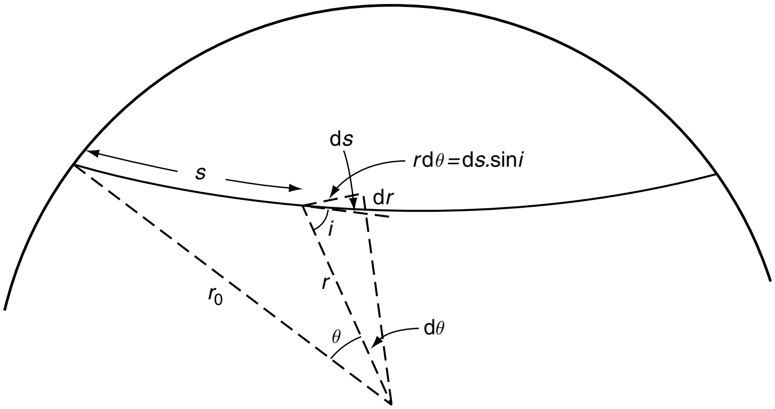


FIGURE 17.9 P- and S-wave speeds, V_P and V_S, in the ak135 model of the Earth developed from body wave travel times (Kennett *et al.*, 1995). Details of the model can be found at wwwrses.anu.edu.au/seismology/ak135.

FIGURE 17.10 Geometry used to obtain an integral expression for travel distance Δ in terms of seismic ray parameter, *p*.



has applied this method to reflected waves from the upper mantle.

To see how travel times are used to deduce the Earth's velocity structure, consider the geometry of Fig. 17.10. A short element d*s* in the path of a ray makes an angle *i* to the radius at that point, so the contribution to Δ by that element is d*θ*, where

$$r\mathrm{d}\theta = \sin i \ \mathrm{d}s. \tag{17.19}$$

Using this relationship to substitute for sin *i* in the expression for *p* (Eq. (17.13)),

$$p = \frac{r \sin i}{V} = \frac{r}{V}\frac{r\mathrm{d}\theta}{\mathrm{d}s}. \tag{17.20}$$

But the small elementary triangle in Fig. 17.10 gives also

$$(\mathrm{d}s)^2 = (\mathrm{d}r)^2 + (r\mathrm{d}\theta)^2, \tag{17.21}$$

and we can eliminate d*s* from Eqs. (17.20) and (17.21),

$$\left(\frac{r^2\mathrm{d}\theta}{Vp}\right)^2 = \mathrm{d}r^2 + (r\ \mathrm{d}\theta)^2. \tag{17.22}$$

With the introduction of a parameter $\eta = r/V$, we obtain

$$\frac{\mathrm{d}\theta}{\mathrm{d}r}=\frac{p}{r(\eta^2-p^2)^{1/2}}. \tag{17.23}$$

Then, integrating with respect to r from the deepest point of penetration of a ray, at radius r', to the surface, r_0,

$$\frac{1}{2}\Delta=\int_{r'}^{r_0}\frac{p\mathrm{d}r}{r(\eta^2-p^2)^{1/2}}. \tag{17.24}$$

This is Abel's integral equation, giving η, and hence V, as a function of r from observed values of Δ and p.

The solution of Eq. (17.24) in a convenient form, known as the Wiechert–Herglotz formula (e.g. Jeffreys, 1959), is

$$\int_0^{\Delta_1}\cosh^{-1}\left(\frac{p}{p_1}\right)\mathrm{d}\Delta=\pi\ln\left(\frac{r_0}{r_1}\right), \tag{17.25}$$

where p_1 is the value of p at $\Delta=\Delta_1$, for the ray that penetrates to radius r_1, at which the wave speed is V_1. Equation (17.25) can be used for numerical integration over any range for which $p(\Delta)$ is a continuous, monotonic function, and so gives $V(r)$ over this range. Different seismic phases give overlapping ranges, allowing interpolation over the breaks. For example, the major break in P-wave travel times is due to the core shadow zone, and since no direct P-waves 'bottom out' in the outer part of the core, they cannot indicate the velocities there. However, SKS-waves involve much shallower refractions into the core and are of interest in examining a suggestion that there is a gravitationally stable layer at the top of the core (Section 22.7).

The PREM earth model (Appendix F) relies heavily on free oscillation periods, but the models that pre-dated free oscillation observations were based entirely on body wave travel times and were quite close to the currently accepted structure. The modern model, ak135, uses this approach.

17.5 Earth models: density variation in a homogeneous layer

Observations of the speeds of both P- and S-waves, V_P and V_S, Eqs. (16.1) and (16.2), allow μ/ρ and K/ρ to be determined:

$$\phi=K/\rho=V_\mathrm{P}^2-\frac{4}{3}V_\mathrm{S}^2, \tag{17.26}$$

$$\mu/\rho=V_S^2, \tag{17.27}$$

but without further information density, ρ, cannot be estimated independently. Earth models based entirely on body wave data were constructed by assuming large regions, where body wave speeds are smoothly varying, to be chemically and mineralogically homogeneous. Then the density increases with depth, z, by hydrostatic compression,

$$-\frac{\mathrm{d}\rho}{\mathrm{d}r}=\frac{\mathrm{d}\rho}{\mathrm{d}z}=\frac{\mathrm{d}\rho}{\mathrm{d}P}\frac{\mathrm{d}P}{\mathrm{d}z}=\frac{\rho}{K}\rho g=\frac{\rho g}{\phi}, \tag{17.28}$$

where P is the pressure and g is gravity. ϕ is a known function of depth by Eq. (17.26), but g depends on the density profile itself, since

$$g=Gm(r)/r^2, \tag{17.29}$$

where $G=6.674\times10^{-11}\,\mathrm{m^3\,kg^{-1}\,s^{-2}}$ is the gravitational constant and $m(r)$ is the total mass inside radius r,

$$m(r)=\int_0^r 4\pi r^2\rho(r)\mathrm{d}r. \tag{17.30}$$

Equation (17.28) is the Williamson–Adams equation for the density profile in a homogeneous layer. It is commonly referred to as the Adams–Williamson equation, although E. D. Williamson is the first author of the now rather obscure paper in which it first appeared.

Since the compressions and rarefactions of seismic waves are adiabatic, the incompressibility in Eq. (17.26) is the adiabatic value, K_S. Therefore, when this is used in Eq. (17.28), the density variation represented by this equation is that due to adiabatic compression. It gives the density gradient in a homogeneous layer with an adiabatic temperature gradient. The most extensive, apparently homogeneous regions within the Earth, the outer core and the lower mantle, are believed to support temperature gradients that are close to adiabatic, so Eq. (17.28) applies directly to these regions, with a minor proviso concerning granular materials, considered below. If the temperature gradient differs from the adiabatic value by an amount

$$\tau = \frac{\mathrm{d}T}{\mathrm{d}z} - \left(\frac{\mathrm{d}T}{\mathrm{d}z}\right)_{\mathrm{Adiabatic}}, \tag{17.31}$$

then a correction term is applied to Eq. (17.28),

$$\frac{\mathrm{d}\rho}{\mathrm{d}z} = \frac{\rho g}{\phi} - \alpha\rho\tau, \tag{17.32}$$

where α is the volume expansion coefficient. This is Birch's generalization of the Williamson–Adams equation (Problem 17.6). In the lithosphere, where the average excess temperature gradient is of order $10^{-2}\,\mathrm{K\,m^{-1}}$ (10 °C/km) the two terms in Eq. (17.32) are of comparable magnitudes (Problem 19.3). The correction term is believed to be significant also in the layer D″, at the base of the mantle.

Equation (17.28), without the added term in Eq. (17.32), must apply very precisely to the outer core, which is homogeneous and adiabatic. Apart from the lithosphere and the D″ layer, and thermal complications in the phase transition zone (Section 22.5), most of the mantle is believed to be quite close to adiabatic. As we point out below, this is the conclusion to the application of Eq. (17.32) to the lower mantle, but in Section 12.6 we point out that different regions of the mantle, especially the lower mantle, must be convectively cooled sequentially, and not simultaneously, so that some heterogeneity is inevitable.

There are two other effects that complicate the picture in the lower mantle. One is observationally indistinguishable from a departure from adiabatic conditions, and applies to granular materials, such as rocks with minerals having different elasticities. It is a consequence of the difference between relaxed and unrelaxed moduli, as discussed in Section 10.4. The unrelaxed modulus, K_{VRH} in Eq. (10.13), deduced from seismic wave speeds, is slightly higher than the relaxed modulus K_{R} in Eq. (10.12), which governs static compression. For the lower mantle mineral mix the difference is listed in Table 18.1. When Eq. (17.32) is used to infer the departure from an adiabatic gradient, τ, a correction to ϕ is required to convert the unrelaxed modulus observed seismologically to the smaller relaxed modulus. Failure to apply this correction leads to a value of τ that is too small. Over the depth range of the lower mantle this leads to a 100 K underestimate of the temperature difference. A temperature increment of 100 K, in excess of the adiabat across the lower mantle, is inferred if Eq. (17.32) is used with uncorrected PREM data. The correction to the relaxed modulus increases this increment to 200 K (Stacey, 2005).

Another complication that prevents precise fitting of the Williamson–Adams equation to the lower mantle arises from a subtle electronic phase transition in iron ions at lower mantle pressures. As discussed in Section 18.9, this transition is smeared over a wide pressure range, possibly almost the entire lower mantle. As for all pressure-induced transitions, the high-pressure state has higher bulk modulus and density, so that there is a slightly greater, but gradual, increase in density with depth than would be expected for constant phase material and there is an 'anomalous' contribution to d*K*/d*P*. The departure of the lower mantle from the Williamson–Adams equation is nevertheless quite modest, but inferences such as departure from the adiabat must be treated with caution.

17.6 Internal structure of the Earth: the broad picture

The first step in elucidating the structure of the Earth was the identification of the major layers and the depths of the boundaries between them. The radial variation in velocity, obtained from body waves and shown in Fig. 17.9, accomplished this. Additional information is needed to deduce the density variation. Frequencies of numerous modes of free oscillation are well observed and the spheroidal modes involve gravitational restoring forces, so inversion of the free mode spectrum allows the development of a complete model, including the density profile. However, inversion calculations need a starting point and this was provided by models based on body wave data and constrained to match the mass and moment of inertia of the Earth.

The crust is the only part of the Earth for which density can be measured directly. However, the isostatic balance of large surface features, such as mountain ranges and continents, allows the density of the uppermost part of the mantle to be calculated, as considered in Section 9.3, (see Fig. 9.6). The value is close to 3370 kg m^{-3}. Studies of upper mantle-derived rocks confirm that this is a reasonable estimate. It provides a starting point for downward extrapolation of density by Eq. (17.28) or (17.32). This extrapolation can be extended as far as the velocity structure indicates that homogeneity is a reasonable assumption, but in the upper mantle a series of velocity increments intervenes.

If we make the simplifying assumption that below the upper mantle there are only two regions, the lower mantle and the core, and that each of these is homogeneous and adiabatic, then Eq. (17.28) applies through each of them. In this case we need to specify only the density at any one level in each and these densities must be selected so that the total mass and moment of inertia of the model match the values of the Earth (Eq. (7.3)). The use of moment of inertia was a vital step in model calculations in the 1930s and 1940s by K. E. Bullen, who showed that the lower mantle was intrinsically more dense than the uppermost mantle by about 700 kg m^{-3}. He therefore identified the upper mantle velocity increments (Fig. 17.9), that we now explain as due primarily to recognized phase transitions in silicates, with a zone of increasing density. The models developed on this basis were remarkably close to our present ideas on Earth structure, as summarized in Appendix F.

Earth models derived from both free oscillation and body wave data are necessarily more secure than models based on one of these data sets alone. The two types of observation each have advantages and limitations. The free modes give accurate global average properties, but have long wavelengths that cannot distinguish fine details, such as whether boundaries are sharp. Body waves can be used to examine details that are inaccessible to free oscillation studies, including lateral heterogeneities, but do not lead directly to density estimates. In considering free oscillations and body waves together, a wide range of frequencies is involved, from 0.3 mHz for the ${}_0S_2$ mode to more than 1 Hz for high frequency body waves. Over this frequency range dispersion is significant (Section 10.7) and must be allowed for, which means that, although Q has the character of a minor correction factor, it is an essential parameter in detailed Earth models.

Appendix F gives selected details of the most widely used earth model, which is known by its acronym, PREM (Preliminary Reference Earth Model). The density structure is plotted in Fig. 17.11(a), with corresponding values of internal pressure and gravity in Fig. 17.11(b). The seismic velocities do not differ, to an extent that is noticeable on a graph, from those of the ak135 model, referred to in Section 17.4. PREM includes dispersion corresponding to a simple Q structure, inferred from free oscillations, and assumed to be uniform over broad depth ranges and to be frequency-independent. This means dispersion given by Eq. (10.29), which, as suggested in Section 10.7, probably overestimates the dispersion in the Earth because Q shows a general increase with frequency. The details in Appendix F refer to properties at 1 Hz. PREM is a global average model, a reference structure for studies of finer details, including lateral heterogeneities, and for discussions of the physics of the deep interior. Anisotropy of the upper layers and separate continental and oceanic structures are given in the original tabulation (Dziewonski and Anderson, 1981).

17.7 Boundaries and discontinuities

Lateral velocity variations are recognized, but are generally small compared with the radial variations displayed in Fig. 17.9. There is a general increase with depth due to the pressure dependence of elasticity, but with discontinuities in radial structure that must be explained in terms of layers with different compositions and/or crystal structures. Important questions about a boundary between layers that can be answered by seismological studies are

(i) what are the differences in the properties (densities, elasticities) of the materials above and below the boundary?

FIGURE 17.11(a) Profile of density, ρ, for the earth model PREM with corresponding zero pressure, low temperature density, estimated by finite strain theory (Chapter 18).

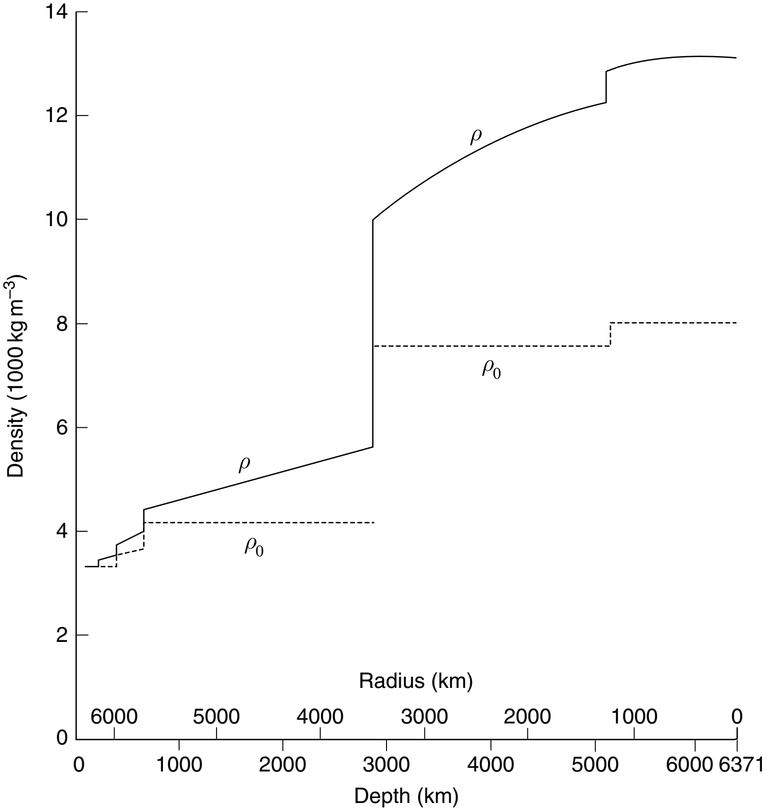
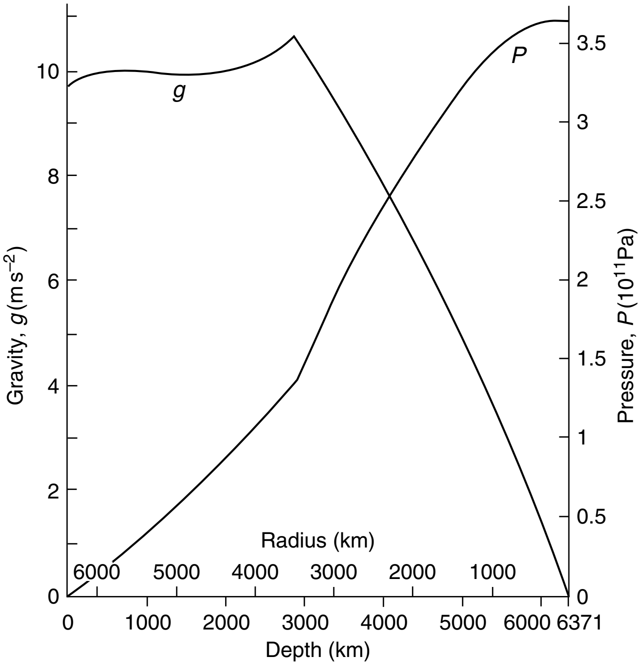


FIGURE 17.11(b) Gravity and pressure corresponding to the density profile in Fig. 17.11(a).



(ii) is the boundary universal or does it occur only in particular areas?
(iii) is it sharp or diffuse?
(iv) is it smooth, undulating or rough?

The contrast in properties across each of the important boundaries in the Earth, as determined in model studies, is unambiguous where the boundaries are sharp and isolated from one another and from graded transitions. A check that these conditions are satisfied is provided by measurements of reflection coefficients and P to SV or SV to P conversions. Reflections occur only where a boundary is sharp, relative to the wavelength used to investigate it. Thus, a boundary that has a reflection coefficient, independent of frequency up to the highest usable frequency, is sharp in the sense of having no observable structure. The core–mantle boundary (CMB) is the most obvious example. This is a boundary between liquid metal and solid, rocky material. However, many reflections from it are very weak because the reflection coefficient is determined by the contrast in acoustic impedance, velocity × density, which for P-waves happens to be nearly the same in the core and mantle (see Problem 16.1, Appendix J). The inner core boundary (ICB) is also found to be sharp (less than 2 km thick according to Kawakatsu, 2006), allowing no more than a thin mushy zone where outer core fluid is freezing out. Reflections observed in the mantle, especially at the 410 and 670 km boundaries, are attributed to solid–solid phase transitions between different mineral assemblages (Tables 2.4a,b). Not all such transitions are sharp. Graded transitions can be observed only by seismic refractions, that is by travel-time studies.

Some boundaries occur at different depths in different areas or are seen in some areas but are missing from others. The boundary between the crust and mantle, the Mohorovičić discontinuity, M layer or Moho, is typically 7 km below the sea floor in oceanic areas, but 35 to 40 km deep under continents and 60 km under major mountain ranges. The different crustal thicknesses are isostatically balanced (Section 9.3) and so the seismological observation of the Moho depth gives a measure of the density contrast between the crust and mantle. Similar undulations of the core–mantle boundary have been sought, but are relatively slight (Doornbos and Hilton, 1989), the strong heterogeneities at the boundary being attributable to the base of the mantle (layer D″). In the case of the CMB the height and typical wavelength of the undulations are important to the mechanical coupling of the mantle to core motions (Section 7.5).

In Section 12.5 the irregularities in D″ are attributed to a structure analogous to continents and ocean basins at the surface (Fig. 12.3). Young and Lay (1990) reported evidence of a boundary at about 250 km above the CMB and it appears possible that this is an upper boundary of a crypto-continent (see Fig. 12.3). This boundary is observed in some areas, but appears to be absent from others, which are therefore assumed to be crypto-oceanic areas where ‘normal’ mantle extends down to the CMB. It will be important to our understanding of the core–mantle interaction to determine what fraction of the CMB is of each kind. The distinction is, in any case, compromised by the transition to the post-perovskite phase near to the base of the mantle. Undulations of this phase boundary would arise from the temperature dependence of the transition pressure.

Recent analysis of digital seismic data have revealed patches of extremely low velocity at the base of the D″ region at the CMB. The origin of these so-called ultra low velocity zones (ULVZs) is controversial, with explanations ranging from temperature effects, possibly accompanied by partial melt, to variations in composition as silicate material interacts with the iron of the core. Maps of ULVZs (Fig. 17.12) correspond roughly to the distribution of hot spots and so ULVZs may be the bases of plumes upwelling from the core–mantle boundary to the surface (Williams *et al.*, 1998). The alternative explanation is that the post-perovskite mineral phase (ppv) at the base of the mantle absorbs iron from the core and iron-rich ppv has low seismic velocities, consistent with ULVZs (Mao *et al.*, 2006).

Topographic maps of the 410 and 660 km discontinuities (Fig. 17.13) show that they undulate by 20 km or more in depth (Flanagan and Shearer, 1998). For the pyrolite model of mantle minerals

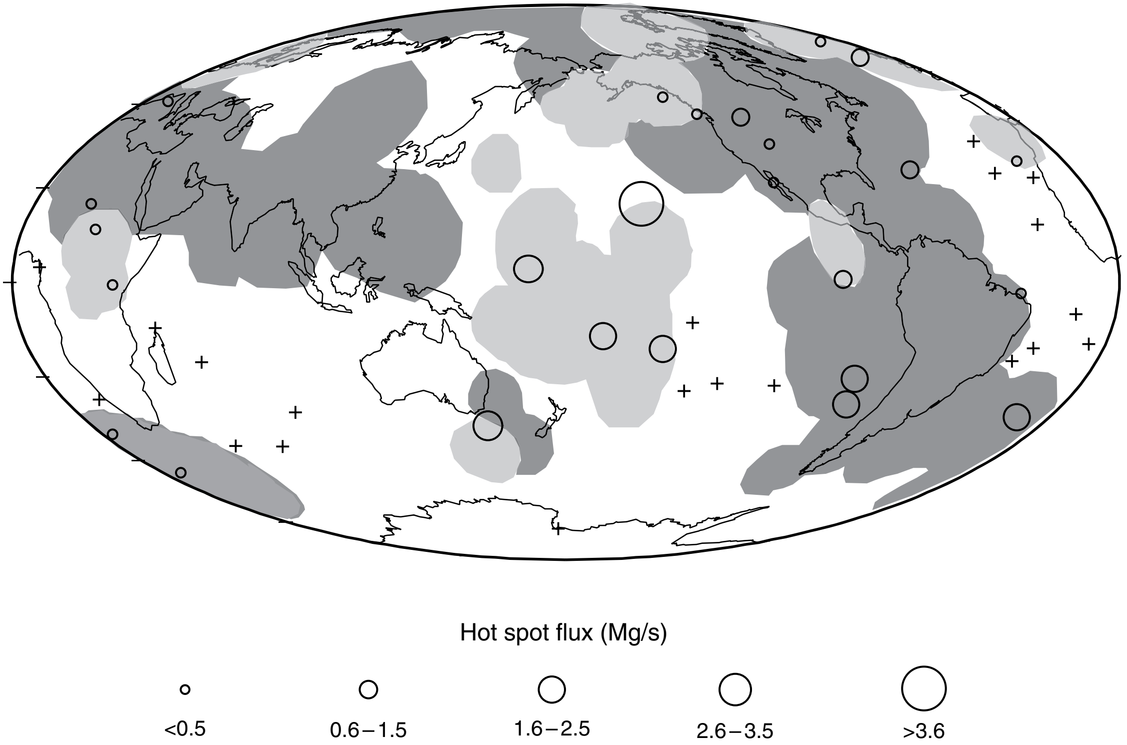


FIGURE 17.12 Locations of ultra low velocity zones (ULVZs) at the base of the mantle. Light shading shows where they are more than 5 km thick, dark shading indicates their absence, or thicknesses less than 5 km. Areas where no determinations have been made are unshaded. Circles represent hot spots with symbol size being proportional to estimated strength in areas studied, and crosses indicating hot spots above regions not yet seismically investigated for ULVZs. (Redrawn from Williams *et al.*, 1998.)

the 410 km boundary marks the $\alpha \rightarrow \beta$ transition of olivine and the 660 km boundary marks the transition to perovskite + magnesiowustite. Clapeyron slopes, $\mathrm{d}T_C/\mathrm{d}P$, for the phase transitions are positive for the 410 km transition and negative for that at 660 km, so that the boundary undulations would be negatively correlated if caused by temperature only. This is not always clear (Flanagan and Shearer, 1998) but evidence of regions exhibiting negative correlation is accumulating. At subduction zones the 660 km boundary is found to be deeper than the global average, but correlated shallowing of the 410 km boundary is not easily observed, perhaps because it is more localized to the slab and presents too limited an area for seismic reflections. Lebedev *et al.* (2002) used seismically estimated variations in transition zone depths and thicknesses, H, beneath stations in Australia and East Asia and temperatures estimated from tomography to obtain Clapeyron slopes. Differential arrival times of P–S converted phases from the 660 km and 410 km discontinuities were converted to H using the tomographic velocities. Temperatures at the discontinuities were inferred by using laboratory estimates of $(\partial \ln V_S/\partial T)_P$ to convert velocity variation, $\delta \ln V_S$, to temperature. A plot of T versus H has a slope $\mathrm{d}T/\mathrm{d}H = 0.13 \pm 0.07$ km/K, in agreement with the laboratory value for pyrolite, 0.13 km/K (Bina and Helffrich, 1994). Variation in transition zone thickness is

$$\delta H = \left(\frac{\mathrm{d}P}{\mathrm{d}T_C}\right)_{660}\left(\frac{\mathrm{d}z}{\mathrm{d}P}\right)_{660}\left(\frac{\partial T}{\partial \ln V_S}\right)_P \delta \ln V_S^{660} - \left(\frac{\mathrm{d}P}{\mathrm{d}T_C}\right)_{410}\left(\frac{\mathrm{d}z}{\mathrm{d}P}\right)_{410}\left(\frac{\partial T}{\partial \ln V_S}\right)_P \delta \ln V_S^{410}. \tag{17.33}$$

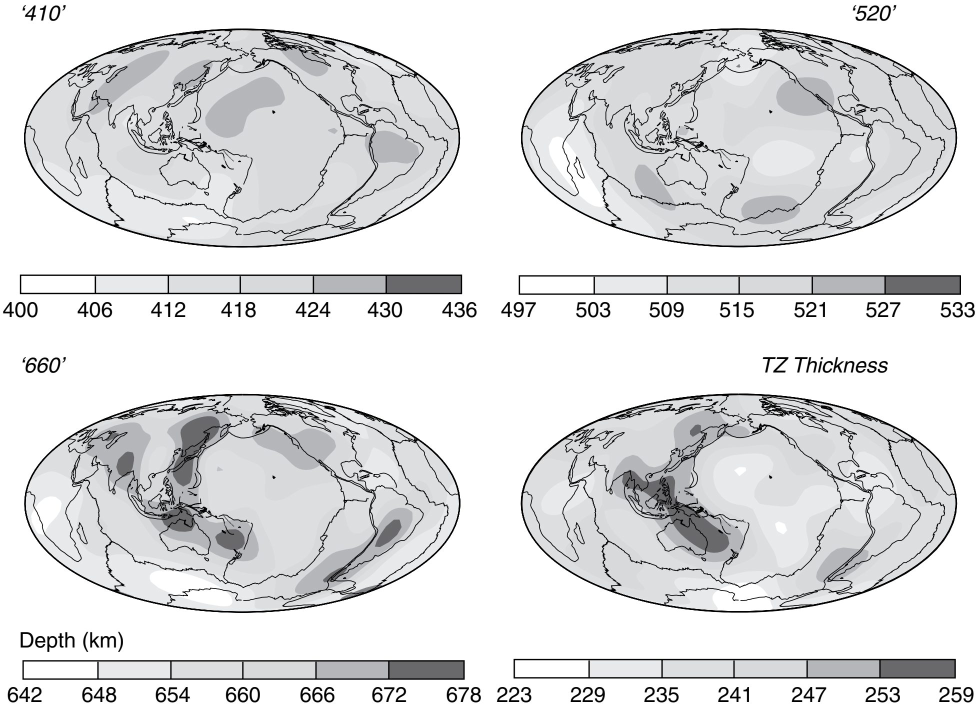


FIGURE 17.13 Undulations of phase boundaries and transition zone thickness (Flanagan and Shearer, 1998).

Data from multiple stations and events were used to estimate $(\mathrm{d}P/\mathrm{d}T_{\mathrm{C}})_{410} = (2 \pm 2\,\mathrm{MPa/K})$ and $(dP/\mathrm{d}T_{\mathrm{C}})_{660} = (-3 \pm 1.5\,\mathrm{MPa/K})$ separately. These values, although imprecise, are compatible with the range of laboratory estimates.

Evidence of an upper mantle boundary 520 km from the surface was seen in stacked digital records of reflected waves by Shearer (1990). This boundary does not appear in PREM. It must occur at a more or less fixed depth to be seen in the globally stacked records, but in regional stacked records this transition has proved to be more elusive. Deuss and Woodhouse (2001) found that in some places the 520 km boundary appears as a single reflector, whereas in others it is either split or not visible at all. Possible candidates are the $\beta - \gamma$ transition in the olivine component and the garnet–perovskite transition in the non-olivine component (Ita and Stixrude, 1992). With variations in temperature or composition these could occur at significantly different depths. A systematic relationship between splitting and tomographic velocity, which would be expected if temperature effects dominate, has not been observed, indicating that composition is also important. The 520 km phase transition is evidently spread over a greater depth range ($\sim$10 km) than the 410 or 660 km transitions ($\sim$2 km) because it does not reflect short period energy (Benz and Vidale, 1993). Its absence from PREM and from refraction studies (Jones *et al.*, 1992) suggests that it arises from a density increment (but not a sharp one), with little velocity change, giving an impedance mismatch and consequent reflections of long-period waves at near-vertical incidence, but little indication in refractions or short-period observations (Rigden *et al.*, 1991; Vidale, 2001).

The boundary at about 200 km (180 km in PREM) was not seen in Shearer's (1990) analyses. These discrepancies suggest fundamental differences between the boundaries, their dependences on temperature and composition and between methods of observation. The boundary at 180 km in PREM is the base of a low-velocity zone, which we identify with the asthenosphere or soft layer of the upper mantle. It is less well developed under continents than under oceans, and if it occurs at a variable depth, when seismograms from different wave paths are added in the stacking process, evidence for a discontinuity is washed out.

17.8 Lateral heterogeneity: seismic tomography

Lateral heterogeneities are most obvious in the crust and occur throughout the mantle. As pointed out in Section 17.1, the outer core must be very homogeneous. It is difficult to imagine significant compositional heterogeneity in the inner core, but its anisotropy is probably quite variable; in any case, seismology gives rather poor resolution. Lateral structure is explored in various ways, especially by seismic tomography. Tomography is a word borrowed from the medical imaging of anatomical structure by multi-directional X-rays, but the analogy is imperfect because X-ray imaging depends entirely on variations in absorption, with no refraction or diffraction. Seismic tomography relies on variations in wave velocity, so that body waves arriving early (or late), relative to waves through a reference model such as PREM, have traversed faster (or slower) paths than average. Travel times over numerous paths in various directions allow the heterogeneities to be localized. But the resolution is limited by the neglect of diffraction (Doornbos, 1992).

The propagation of seismic waves is, in many respects, analogous to the propagation of light waves. Although the similarity includes diffraction, there is an important difference between the way we view diffracted light and the treatment of most seismic signals. When we see an optical diffraction pattern we are looking at the effect of a continuous (usually monochromatic) wave. Light arriving at any point in the field of view is, in general, a phased sum of the contributions from several paths of various optical lengths. In travel-time tomography we consider only the first arrivingpulse, which means that we use only the component of the signal that travels by the fastest path. As mentioned in Sections 16.1 and 16.3, Wielandt (1987) drew attention to problems that can arise in the identification of low- or high-velocity regions from the arrival times of seismic pulses propagating through or diffracted around them.

Consider first a low-velocity sphere. Beyond a limited distance, determined by the size and velocity contrast, the wave diffracted around it arrives earlier than the wave directly through it and the anomaly becomes seismically invisible. For a sphere of high-velocity material, the direct transmitted wave arrives first and the sphere remains visible to much greater distances. Refraction by the sphere spreads out the transmitted early wavefront, so that eventually it is lost, but at intermediate distances the wave spreading has the effect of making the anomalous volume appear larger than it is. Thus, body wave travel times lead to overestimates of high-velocity anomalies and underestimates of low-velocity anomalies. Tomography can only identify broad features in the deep mantle. Some of the things that would be of interest, especially small low-velocity features, such as the stems of ascending plumes, are inaccessible. High-velocity features, such as fragments of subducted slabs, are more visible. On the present evidence it is difficult to judge the importance of the tendency for globally averaged body wave speeds to be biased high.

Recognizing these limitations, travel-time tomography, using high-frequency body waves, has been applied to large-scale features in the lower mantle (Dziewonski, 1984), yielding anomalies with velocity contrasts up to 1%. The spatial pattern was represented by harmonic degrees up to 6, sufficient to outline features about 2000 km across. Subsequent analyses confirmed that at least the larger-scale features are robust, although they are regional averages of structures with unseen finer details. Studies of upper mantle tomography (Woodhouse and Dziewonski, 1984; Zhang and Tanimoto, 1991, 1992) used surface waves with numerous interlocking paths. The principle is similar except that, for any wave path, the different depths are sampled by comparing different frequency ranges (Section 16.5). Subsequent to these early studies, numerous global tomographic models have been presented (e.g., Grand *et al.*, 1997; Grand, 2001; Su and Dziewonski, 1997; Ritsema *et al.*, 1999 – see Fig. 17.14(a); Masters *et al.*, 2000 – see Figs. 17.14(b)–(d)) for which there has been a general convergence on locations and sizes of

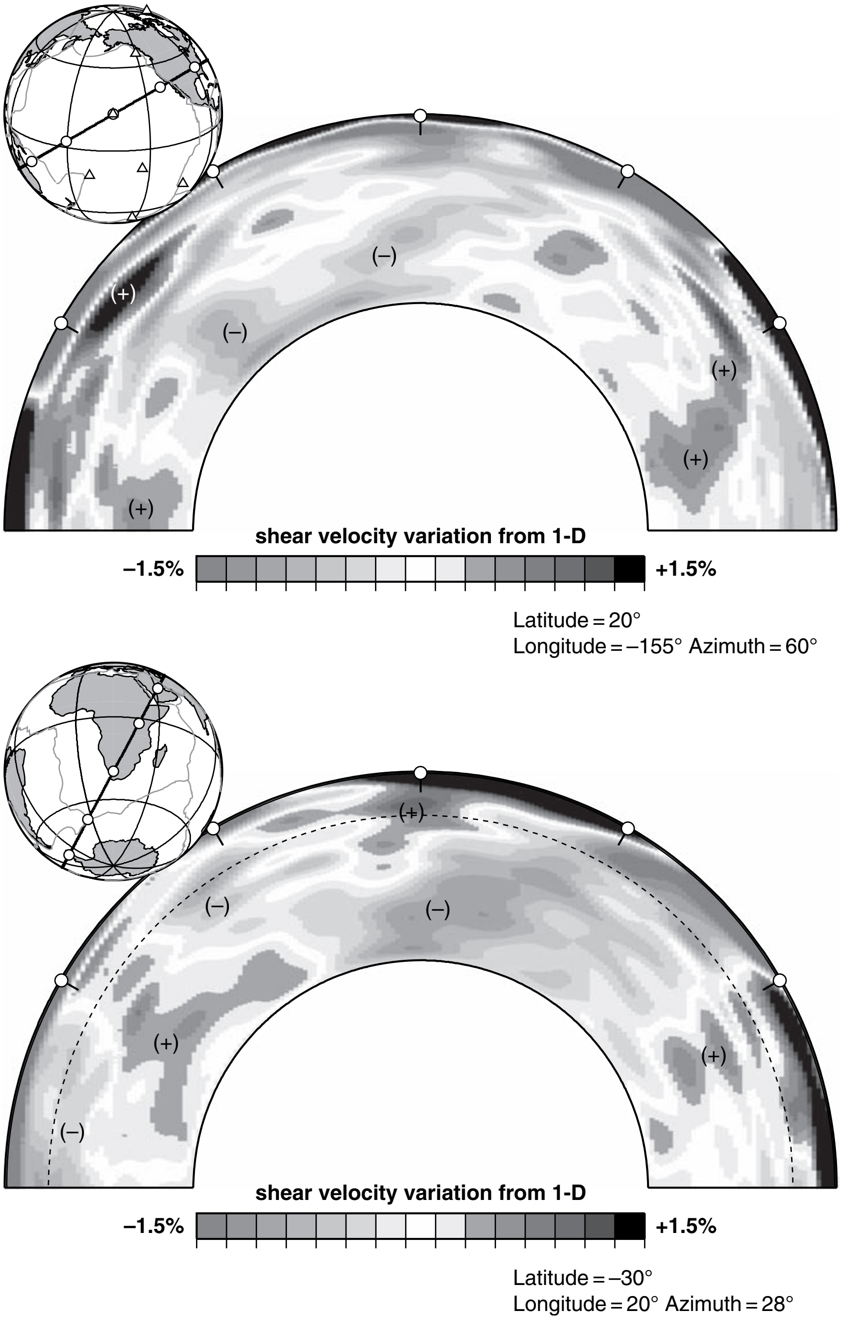
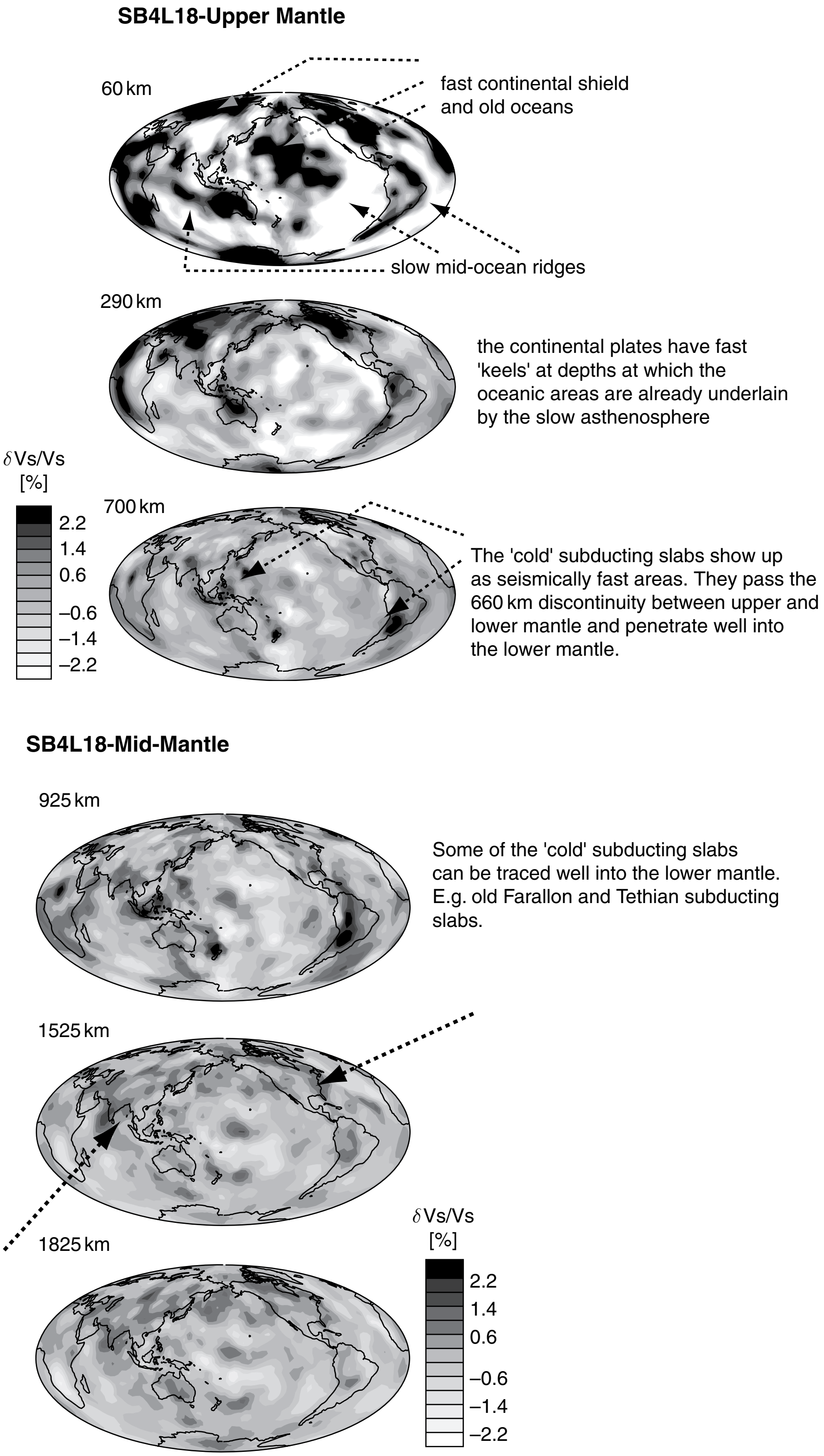

(–)
(+)
(–)
(+)
(+)
(+)
shear velocity variation from 1-D
–1.5%
+1.5%
Latitude = 20°
Longitude = –155° Azimuth = 60°
(+)
(–)
(–)
(+)
(+)
(–)
shear velocity variation from 1-D
–1.5%
+1.5%
Latitude = –30°
Longitude = 20° Azimuth = 28°

FIGURE 17.14(a) Cross-sections of a seismic tomography model (Ritsema *et al.*, 1999), showing Pacific and African superplumes and deep subduction of the Farallon plate under the Americas.

mantle heterogeneities. Of particular note is the discovery that some parts of the subducting slabs beneath North and South America can be traced through the mantle transition zones all the way to the core (Fig 17.14(a)), effectively ending a debate about isolation of the convective circulations of the upper and lower mantles. Some slabs in the western Pacific and beneath Asia appear to pond



SB4L18-Upper Mantle

fast continental shield and old oceans

slow mid-ocean ridges

the continental plates have fast 'keels' at depths at which the oceanic areas are already underlain by the slow asthenosphere

The 'cold' subducting slabs show up as seismically fast areas. They pass the 660 km discontinuity between upper and lower mantle and penetrate well into the lower mantle.

SB4L18-Mid-Mantle

Some of the 'cold' subducting slabs can be traced well into the lower mantle. E.g. old Farallon and Tethian subducting slabs.

FIGURE 17.14 (b), (c) Map views of mantle tomograms from Masters *et al.* (2000). Figures provided by Gabe Laske

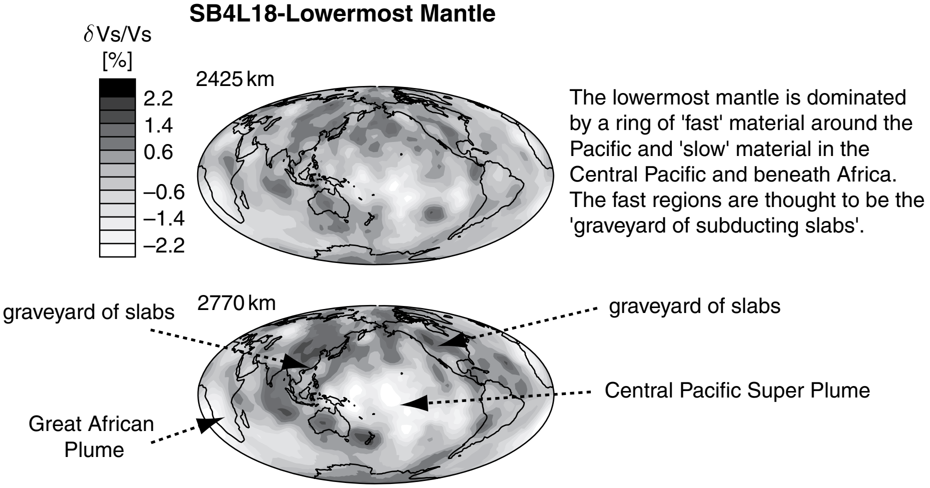

SB4L18-Lowermost Mantle
δVs/Vs
[%]
2.2
1.4
0.6
–0.6
–1.4
–2.2
2425 km
The lowermost mantle is dominated by a ring of 'fast' material around the Pacific and 'slow' material in the Central Pacific and beneath Africa. The fast regions are thought to be the 'graveyard of subducting slabs'.
2770 km
graveyard of slabs
graveyard of slabs
Central Pacific Super Plume
Great African Plume

FIGURE 17.14(d) (Cont.)

in the upper mantle without penetrating the 660 km discontinuity, but deeper high-velocity anomalies suggest they have penetrated in the past and that slab penetration may be episodic.

High velocities are observed beneath cratons (the ancient continental cores), and low velocities below mid-ocean ridges. Localized low-velocity anomalies are observed under hot spots to depths of several hundred km, but at greater depth the picture is more blurred, probably because of a combination of lack of resolution, diffractive healing effects, and reduction of the temperature sensitivity of P-wave velocity with depth (Karato, 1993; Stacey and Davis, 2004, Table 4). Two major low-velocity regions extending from the core into the upper mantle are identified (Fig 17.14(a)), one beneath the Pacific and one beneath Africa, and to the south-west of it. They are referred to as the ‘Pacific and Africa super-plumes’, although whether they actually represent upwelling material or static velocity contrasts is debated. In that they are surrounded by regions that are inferred to have been fed by cold subducting material in the past, they may just represent hotter than average mantle. Alternatively, it has been argued that topographic bulges above the super-plumes are sustained by the dynamics of upward flow. With improved tomographic resolution, it appears that the Pacific super-plume may consist of several plumes (Schubert *et al.*, 2004).

In addition to velocity tomograms, attenuation tomography gives variations in Q in the mantle (Gung and Romanowicz, 2004). Difficulties with both observation and interpretation of Q are discussed in Section 10.5. A general aim of tomographic imaging is to identify features with current or past tectonics, but this can be done only tentatively with the detail available so far.

Generally, tomography has been based on ray theory, which neglects diffraction effects. We point out in Section 16.3 that, to take diffraction into account, we must integrate over the Fresnel volume sampled by the elementary Huygens wavelets that make up a seismic wave. The travel time of a seismic wave is given by ray theory as

$$T = \int_S \frac{\mathrm{d}S}{V(S)}, \tag{17.34}$$

where S is distance measured along the ray and V is velocity. Given a series of travel times, linking various sources and receivers, Eq. (17.34) can be used in a linear inversion to determine the distribution of V^{-1}. As discussed in Section 16.3, ray theory is an infinite-frequency approximation that applies to anomalies of dimension, a, greater than the first Fresnel zone, $a > (\lambda L)^{1/2}$, where L is distance travelled and λ is wavelength (Eq. 16.14). For finite frequencies the relative sizes of the anomaly and the Fresnel zone must be taken into account. In contrast to the simple, essentially one-dimensional expression of ray theory (Eq. 17.34), the travel time for a finite-frequency wave is given by an integral over the whole volume,

$$T = \int_V 1/V(\mathbf{r})K(\mathbf{r},\omega)\mathrm{d}V, \tag{17.35}$$

where $V(\mathbf{r})$ is the velocity distribution, K is a kernel function that formally takes into account the Fresnel volume effects described in Section 16.3. Dahlen *et al.* (2000) give the mathematical detail for computation of K. Equation (17.35) recognizes that objects smaller than the Fresnel zone have negligible effect on travel time, and that diffractive healing diminishes travel time perturbations by features of comparable size. Inversion of travel times using Eq. (17.35), to obtain $V(\mathbf{r})$, effectively reverses diffractive healing and leads to tomographic anomalies that can be up to 60% stronger than those obtained from ray theory tomography (Montelli *et al.*, 2004).

Large features, such as subducting slabs, are well imaged by ray tomography, but smaller more localized anomalies, associated with plumes at depth, are better resolved using finite-frequency tomography. In a recent tomographic inversion by this method, Montelli *et al.* (2004) presented evidence for six plumes that extend from the lowermost mantle, as well as several examples of plumes apparently confined to the upper mantle. Plumes are found to have apparent diameters of several hundred kilometres. These are likely to be the thermal halos of plumes with much narrower rapidly flowing conduits where viscosity is lowest (Loper and Stacey, 1983). The plume viscosity contrast is inferred to be about 10^4:1. In the Montelli *et al.* images the Pacific superplume appears as several sub-plumes.

Interest in tomography arises primarily from its usefulness as an indicator of mantle convection and tectonics, present and past, but the interpretation is not unique. Temperature variations of the seismic velocities are derived from thermodynamic arguments in Section 19.7 (see especially Eqs. (19.70), (19.71) and (19.73)), with numerical values for the lower mantle listed in Table 19.1. The ratio of the temperature sensitivities of V_S and V_P has attracted particular attention. $(\partial \ln V_S/\partial \ln V_P)_P$ depends quite strongly on pressure, increasing from 1.7 to 2.5 over the lower mantle range (column 4 of table 19.1), and this is in general agreement with tomographic observations of V_P and V_S by Robertson and Woodhouse (1996a) and Su and Dziewonski (1997), indicating that the variations have a thermal explanation. However, closer examination shows that this is not a sufficient explanation. These data sets agree that when presented as a comparison of S-wave anomalies with variations in bulk sound velocity, V_ϕ, given by Eq. (19.72), they are seen to be negatively correlated below about 2000 km depth or perhaps uncorrelated (Kennett *et al.*, 1998). The significance of V_ϕ is that it does not depend on the rigidity modulus, μ, but only on the bulk modulus K_S (and ρ). The negative correlation, that is a decrease in V_ϕ corresponding to an increase in V_S and vice versa, can be interpreted as a thermal effect only by supposing that K_S decreases with temperature more than does ρ, and by Eq. (19.73) that means $\delta_S < 1$. δ_S is calculated by Eq. (19.64), a thermodynamic identity, in terms of parameters that are constrained by the equation of state of the lower mantle (as detailed in Stacey and Davis, 2004 and listed in Appendices F and G), by which δ_S remains greater than unity throughout the mantle (see Section 19.7). We must therefore appeal to compositional variations in addition to the thermal effects.

Forte and Mitrovica (2001) considered the compositional variations of the lower mantle in terms of the SiO_2–FeO–MgO system with varying proportions of these components. They used seismic and modelling data to infer velocity changes due to changes in the molar fraction of iron, $X_{\mathrm{Fe}} = [\mathrm{FeO}]/([\mathrm{FeO}] + [\mathrm{MgO}])$, and the molar fraction of perovskite $X_{\mathrm{Pv}} = [\mathrm{Pv}]/([\mathrm{Pv}] + [\mathrm{Mw}])$, where Pv refers to perovskite and Mw refers to magnesiowustite. Then

$$\begin{aligned}\delta \ln V_S &= \frac{\partial \ln V_S}{\partial T}\delta T + \frac{\partial \ln V_S}{\partial X_{\mathrm{Pv}}}\delta X_{\mathrm{Pv}} + \frac{\partial \ln V_S}{\partial X_{\mathrm{Fe}}}\delta X_{\mathrm{Fe}},\\ \delta \ln V_\phi &= \frac{\partial \ln V_\phi}{\partial T}\delta T + \frac{\partial \ln V_\phi}{\partial X_{\mathrm{Pv}}}\delta X_{\mathrm{Pv}} + \frac{\partial \ln V_\phi}{\partial X_{\mathrm{Fe}}}\delta X_{\mathrm{Fe}}.\end{aligned} \tag{17.36}$$

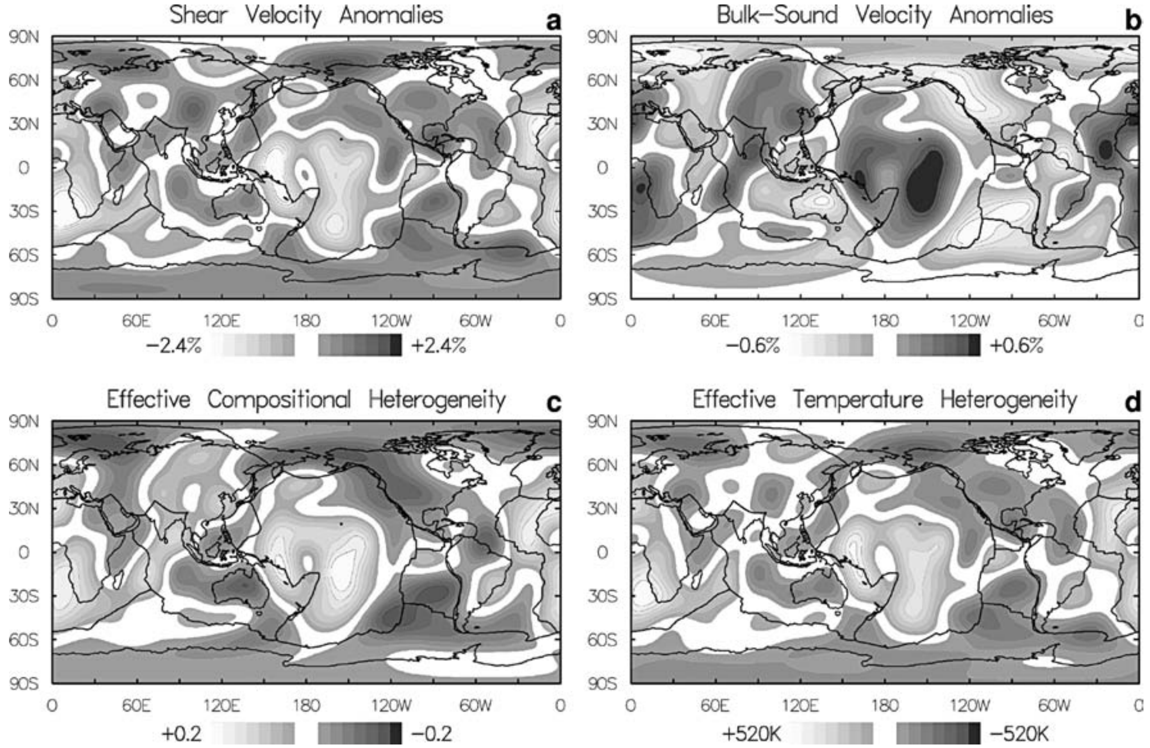


FIGURE 17.15 Tomographic shear and bulk sound velocity models compared with inferred temperature and compositional variation at a radial distance of 3600 km (about 120 km above D″). (Forte and Mitrovica, 2001.)

With values of temperature derivatives at radius 3600 km, 120 km above the core– mantle boundary and the Forte–Mitrovica estimates of compositional derivatives, these equations become

$$\delta \ln V_S = -2.5 \times 10^{-5}\delta T - 0.16 \times \delta X_{Pv} - 0.22 \times \delta X_{Fe},$$
$$\delta \ln V_\phi = -0.12 \times 10^{-5}\delta T + 0.045 \times \delta X_{Pv} + 0.048 \times \delta X_{Fe}. \quad (17.37)$$

On this basis Forte and Mitrovica (2001) mapped δX_{Fe} with maximum values of about ± 0.01 (Fig. 17.15). Assuming temperature variations of about 100 K, temperature has the dominant effect on V_S in Eq. (17. 37) but V_ϕ is most affected by compositional variations. This can account for the negative correlation between V_ϕ and V_S. If this is the explanation, then high temperatures are correlated with high iron content, so that the thermal and compositional effects on density are, at least partly, compensating.

17.9 Seismic anisotropy

With rare exceptions, published tomographic inversions have been based on the assumption that that the velocity anomalies are isotropic. This has been more a matter of necessity than a fundamental requirement, because the data are insufficient to resolve the extra parameters required to describe anisotropy, but the pervasiveness of anisotropy in the crust and upper mantle suggests that interpretations based on isotropy could be simplistic. Anisotropy is sought in terms of variations in travel times of seismic waves with direction and polarization and may be due to alignment of intrinsically anisotropic minerals or fabric made up of aligned elastic heterogeneities. If the fluctuations in anisotropy are sufficiently random the region appears isotropic. Crustal anisotropy is highly variable. On a small (kilometre) scale, velocity variations can be very large (as much as 20%), but on larger scales (tens of kilometres) the anisotropy appears to be almost incoherent. By contrast, the uppermost mantle exhibits long-wavelength coherent anisotropy, probably caused by alignment of olivine crystals. Olivine is highly anisotropic with orthorhombic symmetry (see Tables 10.2a, b). When subjected to finite shear deformation, fast directions align in the direction of extension, and, in the case of extreme shear, rotate to become parallel to the shear plane itself.

A complete description of the effects of mantle fabric on seismic wave propagation requires a full anisotropic tensor (Crampin, 1977), but determination of all 21 components of the elastic tensor (Chapter 10) is not feasible. Typically, several parameters are obtained from a particular set of observations while the remaining coefficients are constrained to be averaged isotropic values. The resulting anisotropy parameters are found to be in good agreement with laboratory measurements on mantle rocks, such as ophiolites and xenoliths, which typically exhibit P-wave anisotropy of about 8% and S-wave anisotropy of 4% (Long and Christensen, 2000).

The reference Earth model (PREM, Dziewonski and Anderson, 1981) describes the uppermost 195 km of the mantle as an anisotropic solid with cylindrical symmetry about a vertical axis. This is referred to as transverse isotropy, and requires five depth-dependent elastic moduli. The remainder of the PREM mantle is isotropic with two moduli. Transverse isotropy was used to model the observation that velocities of mantle Love waves are higher than those of mantle Rayleigh waves. Azimuthal anisotropy has been recognized in oceanic mantle since the 1960s with the observations that P_n-waves, critically refracted in the mantle beneath the crust, travel faster in the spreading direction than perpendicular to it (Hess, 1964; Raitt *et al.*, 1969; Keen and Barrett, 1971). In young ocean lithosphere Rayleigh wave fast directions are also parallel to plate motion direction (Forsyth, 1975; Nishimura and Forsyth, 1989; Montagner and Tanimoto, 1991; Laske and Masters, 1998). However, in older sea floor, Nishimura and Forsyth (1989) found that the effect weakens, possibly due to over-printing of different directions of absolute plate motion. Nishimura and Forsyth analysed the spatial and depth dependence of anisotropy of the Pacific Ocean floor, finding it to be confined to the upper 200 km. Thus, the anisotropy occurs mainly in the mantle lithosphere, but extends into the underlying asthenosphere.

Birefringent effects on SKS waves are used to determine azimuthal anisotropy. Because the SKS wave is generated by a conversion at the CMB of a core-travelling P-wave, the polarization of the emerging S is SV. When the S-wave encounters azimuthally anisotropic material it splits into fast and slow components that arrive at the surface out of phase, making the motion elliptical. S-wave polarizations are termed radial when parallel to the great circle between source and receiver, and transverse if normal to it. Knowing that the initial polarization is radial, it is possible to reconstruct the incident waveform to determine the fast direction and the phase delay. The splitting process effectively differentiates the waveform such that the transverse component is the time derivative of the incident radial component. To see this, consider a nearly vertically incident split shear wave. Let the radial component of the polarization be $f(t)$ and suppose the fast direction makes an angle ϕ with the radial component, the slow direction an angle of $\phi + \pi/2$ and that the travel time difference is δt. After splitting, the fast component is $f(t + \delta t/2)\cos\phi$ and the slow component is $-f(t - \delta t/2)\sin\phi$. Resolving these into the transverse direction (i.e. at $\pi/2$ to the radial direction), one obtains

$$\begin{aligned}\text{transverse} &= [f(t + \delta t/2) - f(t - \delta t/2)]1/2\sin 2\phi \\ &\approx (1/2)\delta t \sin 2\phi \frac{\mathrm{d}f(t)}{\mathrm{d}t}, \text{ for } \delta t \text{ small.}\end{aligned} \tag{17.38}$$

This property that the transverse component is 90° out of phase with the radial component is useful for recognizing split energy in the presence of interfering phases, such as S, that normally have radial and transverse components in phase.

Numerous studies of SKS splitting have been published (e.g., Silver, 1996). Typically phase delays are 1 to 2 s, corresponding to 2 to 4% azimuthal anisotropy in the uppermost mantle

(200 km depth). Fast directions are often aligned with finite extensions of this layer, but there are many exceptions that require explanation. Montagner *et al.* (2000) found that splitting predicted from surface wave anisotropy is in good agreement with observations for regions undergoing large-scale coherent tectonics, such as the western United States and Central Asia.

Seismic anisotropy in the upper mantle is thought to be due mainly to oriented olivine crystals. Since olivine crystals are orthorhombic, a complete description requires nine independent elastic constants and three Euler angles to define orientation. By combining different measures of anisotropy such as SKS splitting, azimuthal variation of surface wave dispersion, and misalignment of wave motion relative to the directions from which waves arrive, it is possible to infer depth-averaged values of these constants (e.g. Davis, 2003).

At depths greater than 220 km the mantle appears to be less anisotropic, but is unlikely to be perfectly isotropic (Boschi and Dziewonski, 2000; Panning and Romanowicz, 2006). The alignment in the uppermost mantle is thought to be caused by dislocation creep with the different crystalline planes of olivine having different effective viscosities and resulting in deformation that depends on the orientations of individual crystals. This causes an alignment of crystal axes as strain progresses. At high homologous temperatures (T/T_M), the dominant deformation mechanism is believed to be diffusion creep, which destroys an existing fabric without generating a new one, and this is consistent with weaker anisotropy of the asthenosphere. The lower mantle is less anisotropic than the upper mantle, probably because the minerals are less anisotropic. There is some evidence of anisotropic travel times for waves grazing the core–mantle boundary, which could be explained as a fabric in the D″ layer.

The inner core is anisotropic. Compressional waves travelling through it parallel to the rotation axis have travel times 3 s to 4 s shorter than waves travelling in the equatorial plane (Poupinet *et al.*, 1983). Explanations in terms of axial elongation of the inner core were discounted when it was shown that the splitting of certain free-oscillation frequencies was explained by an anisotropic inner core (Woodhouse *et al.*, 1986; Tromp, 1993). Later body wave analyses have indicated lateral variations in the anisotropy of the inner core (Creager, 1997). The anisotropy is an important clue to the mechanism of inner core formation by accretion primarily on its equator and steady deformation towards equilibrium ellipticity (Yoshida *et al.*, 1996). It is thought that the inner core consists of ε-iron, a hexagonally close-packed phase with cylindrical symmetry, and that anisotropy of the inner core is an expression of its crystalline alignment.

Inner core anisotropy is not perfectly aligned with the rotation axis and is highly variable. Song and Richards (1996) recognized the possibility of seeing differential rotation of the inner core, relative to the mantle, from slow variations of travel times for seismic waves with inner core paths. This has important implications for core physics and the dynamo, and is discussed in Section 24.6. The original report was followed by widely different estimates of rotation rate as well as refutations, but recent better controlled observations by Zhang *et al.* (2005), illustrated in Fig. 17.16, leave little doubt that there is a real effect. Interpretation depends on imprecisely observed inner core structure. It should also be noted that electromagnetic coupling of the inner core to the complicated, irregular field of the outer core allows the possibility that the inner core rotation axis differs by a few hundredths of a degree from that of the mantle. The observations may not see a simple differential rotation about a common axis, but polar wander of the inner core.

The misalignment of the axis of the inner core anisotropy invites comparison with the misalignment of the magnetic dipole axis. In both cases we appeal to Curie's (1894) principle of symmetry, according to which no effect can have lower symmetry than the combination of its causes. Paterson and Weiss (1961) discussed geological applications of this principle. In these situations the principle must be applied statistically, that is we must consider the long-term average alignment of the anisotropy axis to be constrained by Curie's principle and not the instantaneous alignment, just as we average

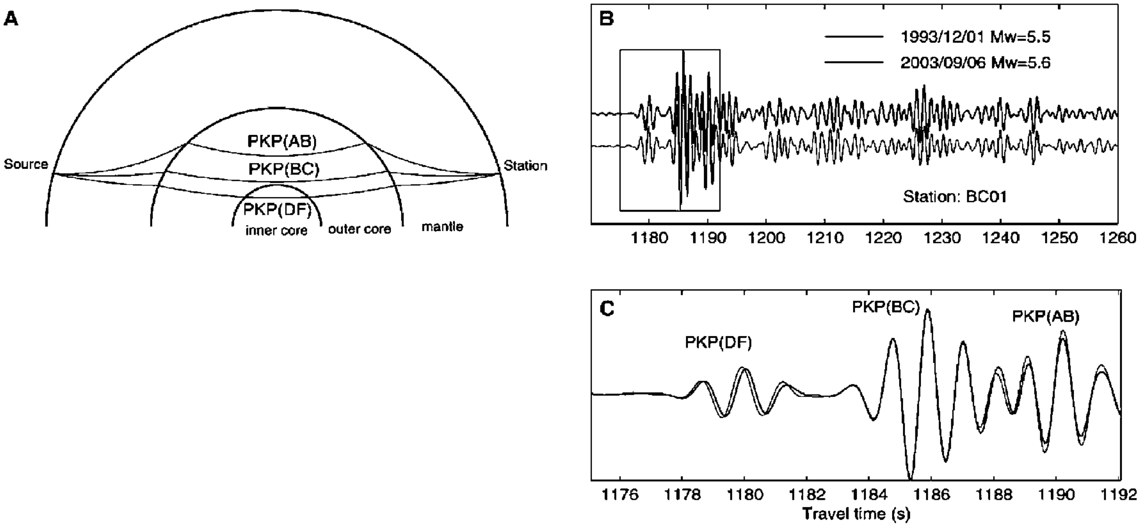
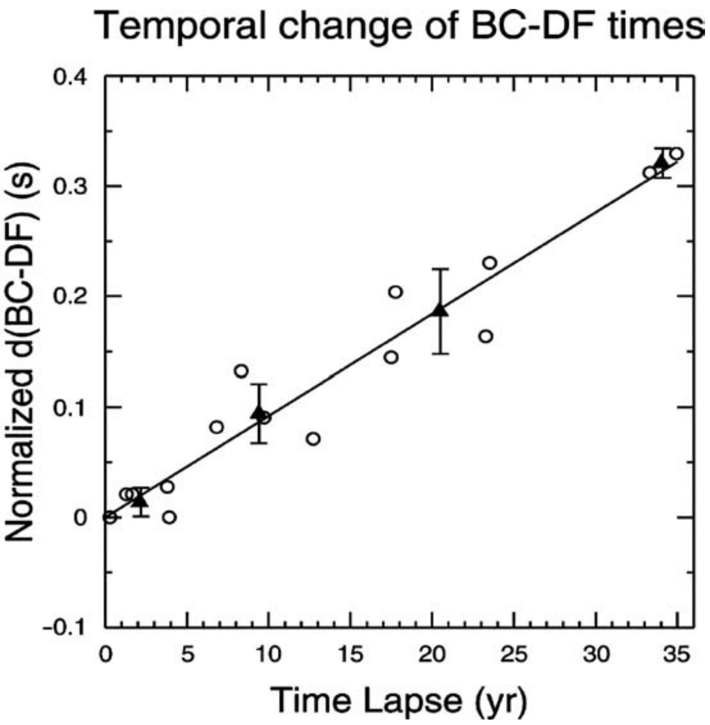


FIGURE 17.16(a) Ray paths of PKP waves and an example of a waveform doublet used to detect a temporal change of travel times through the inner core (Zhang *et al.*, 2005). A doublet refers to a pair of earthquakes separated in time, with highly correlated waveforms suggesting that they have nearly identical locations. The phase change between waves that pass through the inner and outer cores increases with the time interval between events. This has been taken as evidence that the inner core is rotating, relative to the mantle.

FIGURE 17.16(b) Difference of BC–DF times, d(BC–DF) (Fig. 17.16(a)) at station COL (College, Alaska) as a function of the time separation between the two events of a doublet. (Zhang *et al.*, 2005.)



the magnetic field over thousands of years to demonstrate the axial dipole principle. We refer to this principle in Sections 24.5 and 25.3. On a similar time scale (controlled by core motions) the inner core anisotropy must, by Curie's principle, be aligned with the rotation axis. Two effects must be expected from the electromagnetic coupling to irregular motion in the outer core. The material of the inner core may move relative to its own rotation axis (polar wander of the inner core) and the axis may be misaligned with that of the mantle. By Curie's principle, the second of these effects would be averaged out over thousands of years, but the first one would cause permanent heterogeneity of grain alignment in the inner core. It appears possible that seismic observations may eventually clarify the details of present inner core rotation.

18

Finite strain and high-pressure equations of state

18.1 Preamble

An equation of state is a relationship between volume, V, pressure, P, and temperature, T, of a specified mass of material. If density, ρ, is used instead of V then specification of mass is unnecessary, and in some treatments V is specific volume, $1/\rho$, but in this text V is the volume of an arbitrary mass, m. Another alternative is molar volume, but moles can be inconvenient units for materials with non-integral proportions of different elements. The simplest and most familiar equation of state is the ideal gas equation for n moles of gas, $PV = nRT$, where $R = N_A k = 8.314\,47\,\mathrm{J\,mol^{-1}\,K^{-1}}$ is the gas constant and is related to Boltzmann's constant, $k = 1.380\,65 \times 10^{-23}\,\mathrm{JK^{-1}}$ by Avogadro's number, $N_A = 6.022\,14 \times 10^{23}\,\mathrm{mol^{-1}}$. It is an example of a complete equation of state, by which is meant one representing V as a function of both P and T, not that it gives all of the properties. In dealing with condensed matter (solids and liquids) it is generally convenient to consider the P and T effects separately. Compression at constant T is described by the isothermal bulk modulus or incompressibility, $K_T = -V(\partial P/\partial V)_T$. This is a parameter of elasticity theory (Chapter 10), which is restricted to small strains, that is, for volume compression, $P/K_T = -\Delta V/V \ll 1$. We refer to elasticity theory as infinitesimal strain theory, in which K_T is treated as constant. For stronger compressions we need a finite strain theory, in which K_T varies with P, sometimes referred to as an isothermal equation of state. For a complete equation of state it must be supplemented by information about thermal expansion.

In the Earth, bulk modulus varies by a factor exceeding 10 (for core material, relative to zero pressure), making the necessity for a finite strain theory obvious. Unfortunately there is no agreed, general theory, but only rival ideas, all of which must be recognized as empirical. Even the definition of strain itself is problematic. Most geophysical discussions have followed the ideas of Birch (1952), who pioneered the interpretation of seismological observations of the deep Earth in terms of a theory that was developed from the classical elasticity study of Love (1927). Birch's theory is most simply presented in terms of interatomic potentials, although its original development was quite different. It works reasonably well for minerals if used to estimate density as a function of pressure, but if derivatives ($\mathrm{d}K/\mathrm{d}P$, $\mathrm{d}^2K/\mathrm{d}P^2$) are required, as in the calculation of thermal properties (Section 19.3) or for extrapolation outside the range of fitted data, then it is not satisfactory and the use of what we call derivative equations is necessary. But the search for further improvements continues and thermodynamic arguments are central to this search.

A finite strain theory makes a particular assumption about the variation in temperature, most commonly $T =$ constant or even $T = 0$, but other alternatives are possible. For geophysical purposes it is often more convenient to use an adiabatic finite strain equation, so that no thermal correction is needed to represent $V(P)$ or $P(V)$ within the large regions of the Earth where the temperature gradient is believed to be close to adiabatic. It uses directly the seismologically observed bulk modulus, which is not K_T but K_S, representing adiabatic compression,

$$K_S = -V(\partial P/\partial V)_S = K_T(1 + \gamma\alpha T), \tag{18.1}$$

where

$$\gamma = \alpha K_T/\rho C_V = \alpha K_S/\rho C_P \tag{18.2}$$

is the Grüneisen parameter, a subject of Section 19.3. In the Earth K_S exceeds K_T by 5% to 10%. Another possibility is an equation representing compression on the melting curve, with modulus K_M, which is referred to in Section 19.4.

The thermal properties required to add an arbitrary temperature variation are not independent of the elastic properties described by a finite strain theory. They are controlled by the same atomic forces. The Grüneisen parameter is the essential theoretical link because it can be represented in terms of elastic moduli and their pressure derivatives (Section 19.3). This means that such calculations require derivatives of the finite strain equations on which they are based and care is required in selecting a theory for which the derivatives meet thermodynamic criteria.

Most finite strain equations are relationships for $P(V)$ and cannot be inverted to give analytical expressions for $V(P)$. Partly for this reason it is often preferable to add thermal effects by considering the application of heat at constant V rather than constant P. The increase in P with T at constant V is termed thermal pressure, being the pressure that prevents thermal expansion. We have a thermodynamic identity (Appendix E)

$$(\partial P/\partial T)_V = \alpha K_T, \tag{18.3}$$

so that

$$P_{\mathrm{Th}} = \int_0^T \alpha K_T \mathrm{d}T = \rho \int_0^T \gamma C_V \mathrm{d}T, \tag{18.4}$$

integrated at constant V. It is convenient to the application of Eq. (18.4) that, for insulators at high temperatures, the product (αK_T), like (γC_V) in Eq. (18.4), is only slightly dependent on T (at constant V). For metals, including the Earth's core, an electron heat capacity proportional to T must be added.

Finite strain equations do not carry through phase transitions. Different phases of a material have different physical properties and the parameters of an equation of state are specific to a particular phase or crystal structure. Mineral phase transitions in the mantle are marked by discontinuities in properties such as K_S or K_T and α as well as ρ. However, some mineral phase transitions (those in Table 2.4b) are not sharp but smeared over ranges in pressure, so that seismological estimates of parameters such as $\mathrm{d}K/\mathrm{d}P$ lose their normal meanings over these ranges and equations of state can be applied only very cautiously. They are fully effective only when applied to those parts of the Earth that are uniform in mineral structure as well as composition, the outer core and most of the lower mantle.

The changes in properties such as K_S that occur at phase transitions can be attributed to the changes in atomic coordination. Thus, it is interesting to note that at the inner core boundary (ICB), between solid and liquid iron alloys, the change in properties is very slight. Most of the density difference is due to a difference in the abundance of solutes (Section 2.8), which are nevertheless dilute enough to have little effect on bulk modulus. The very slight change in K_S with melting is characteristic of metals that melt with little change in atomic coordination, as described by the dislocation theory of melting (Section 19.4) and expected at pressures sufficient to ensure that both solid and liquid are close-packed.

Appendix F gives selected details of the Earth model PREM. This is a parameterized model in the sense that density and the seismic wave speeds, V_P and V_S are fitted to polynomials in radius over different ranges, the parameters being the polynomial coefficients. While this is mathematically convenient and represents the variations of ρ and K_S with P (Fig. 18.1) adequately for some purposes, it is of limited use for equation of state studies because the derivatives,

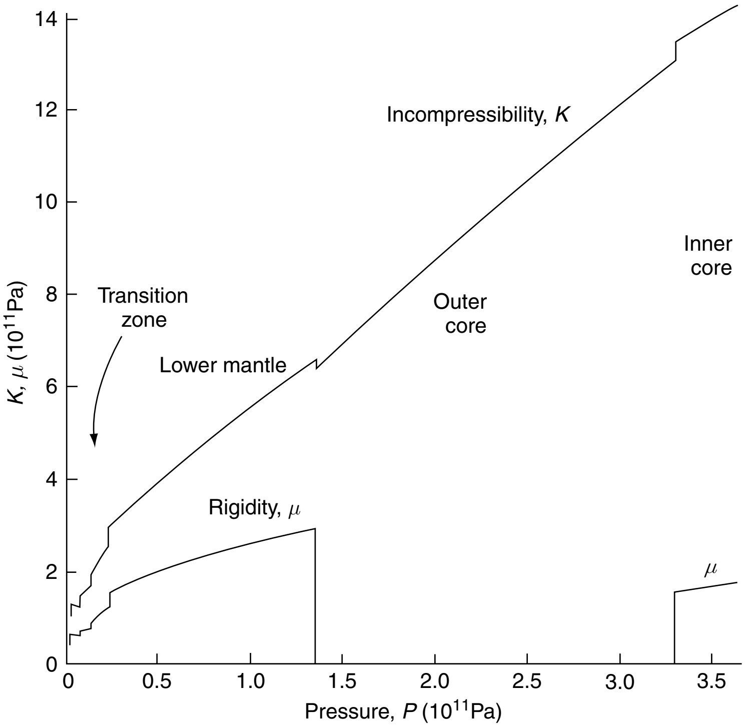


FIGURE 18.1 Variations of the elastic moduli, K and μ, with pressure, P, for the PREM Earth model.

$K'_S = (\partial K_S/\partial P)_S$ and $K''_S = (\partial^2 K_S/\partial P^2)_S$, are incompatible with any plausible equation of state. They show trends that are consequences of the model parameterization and do not follow the necessary monotonic pressure variations. This can be handled by fitting the data to an equation of state, but then, of course, the chosen equation imposes its own functional form and so determines details of the inferred properties. In this circumstance information additional to the fitted data is required to ensure that the equation gives physically real behaviour for properties that depend on derivatives of the equation. Fits of lower mantle and core data to a derivative equation are listed in Appendix F for comparison with the PREM tabulation on which they are based.

As discussed in Section 17.5, in a homogeneous, adiabatic layer the variation in density with pressure is given by $\mathrm{d}\rho/\mathrm{d}P = \rho/K_S = 1/\phi$, which is obtained directly from seismic velocities. This is a semi-independent check on the $P(\rho)$ variation of an Earth model and so provides a test for homogeneity. In layers where Earth models clearly indicate homogeneity, the outer core and most of the lower mantle, there is, therefore, a redundancy in the information that is very useful to equation of state studies. Instead of fitting just a P–ρ equation, we can differentiate it to obtain a K–ρ equation and, by taking the ratio, we have a P/K vs ρ equation that can be fitted. Since both P and K are listed in Earth models, this eliminates the unknown K_0 from an initial data fit, making the fit correspondingly more certain. For this reason, as well as the surety of the pressure scale, over the pressure ranges of the outer core and lower mantle, there is a strong advantage to the use of Earth model data for testing finite strain equations.

18.2 High-pressure experiments and their interpretation

Laboratory simulation of deep Earth conditions of pressure and temperature allows comparison of measured properties of candidate terrestrial materials with seismological data, providing checks on the theories discussed in this chapter as well as the composition and mineral structure of the Earth. The several techniques each have advantages and limitations. The first to be used was hydrostatic compression in fluid-filled pressure vessels, as in early work by P. W. Bridgman. This still has an important role, in spite of more recent developments of methods of attaining much higher pressures, because it allows precise measurements of acoustic velocities as functions of pressure and therefore yields equation of state derivatives, $K_S = (\partial P/\partial \ln \rho)_S$ and $(\partial K_S/\partial P)_T$ at modest pressures. It also offers the surety that pressure is hydrostatic, which is not always straightforward with higher pressure methods. The next derivative, $\mathrm{d}^2K/\mathrm{d}P^2$, either isothermal or adiabatic, is sometimes reported, although pressure calibration is not generally precise enough to make this useful.

Bridgman also experimented with the compression of small, solid samples between tapered anvils of very hard materials (sometimes crossed anvils that required no special care in alignment). The modern use of diamond anvils pursues this idea with great success (Fig. 18.2), thanks to the remarkable properties of diamond. Pressures of a few hundred GPa (a few megabars) have been achieved, although most experiments have been restricted to a range below 100 GPa. The transparency of diamond allows both sample observation and laser heating. Diamond is reasonably transparent also to X-rays, permitting *in situ* measurement of X-ray diffraction by the very small samples and therefore determination of both crystal structure and lattice spacing (and hence density) at very high pressures. The method has proved to be extremely versatile, with care to ensure that pressure is reasonably close to hydrostatic and that, in heating experiments, temperature is uniform as nearly as possible. As with other techniques, pressure calibration is problematic. But the small size and relative simplicity of diamond anvil cells have made them readily accessible and they are used for a wide range of experiments in many laboratories.

An early success in the application of diamond anvils to mineral physics was a series of experiments by Liu (1976), who heated a variety of minerals by lasers to 1000 °C under strong compression, showing that at a pressure corresponding to the 660 km transition in the mantle (23.5 GPa) these minerals were transformed predominantly to an orthorhombic perovskite phase of $(Mg,Fe)SiO_3$. Although such a transition had been anticipated on the basis of similar transitions at lower pressures in chemical analogues, the direct confirmation that it occurred in mantle silicates was a crucial step in our understanding of the lower mantle. It also highlighted the usefulness of diamond anvils in high pressure mineralogy.

Some experiments require much larger specimen sizes than are possible with diamond anvils. Static experiments on larger specimens use multi-anvil presses with sample enclosures of cubic or octahedral shape compressed by six or eight rams driven by hydraulic pistons. Specimens can be surrounded by relatively soft material to ensure a close approximation to hydrostatic compression and heated by internal electric elements from which the massive

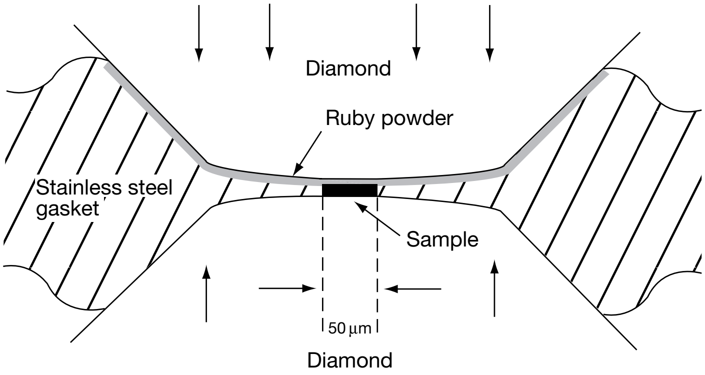


FIGURE 18.2 Geometry of diamond anvil cells in which pressures up to 250 GPa have been reached. Powdered ruby fluoresces at a wavelength that varies with pressure and is used for pressure calibration, although not generally to the highest pressures.

pistons are insulated, with temperature monitored by thermocouple. A particular use of such apparatus is the production of high pressure minerals, notably the perovskite mentioned above, which is believed to be the dominant component of the lower mantle but survives in a metastable state at laboratory pressure and temperature. After crystal growth at high pressure and temperature it is quenched (cooled to laboratory temperature while still at high pressure) before decompression. Some other minerals, including Ca-silicate perovskite, which is almost certainly a minor lower mantle constituent, do not survive decompression and must be observed *in situ* in the high pressure apparatus.

The highest pressures are achieved transiently in shock wave experiments. This method is a by-product of atomic weapons development, and the original application to terrestrial materials was intended, at least partly, as a check on the technique because the densities of these materials at very high pressures were believed to be known. Early results (McQueen and Marsh, 1966; McQueen *et al.*, 1967) confirmed that phase transitions converted familiar minerals to denser forms, consistent with the lower mantle, but that the core could not be explained by further phase transitions and must be interpreted as an alloy dominated by iron, with minor addition(s) of lighter element(s). Observations are made by high speed photography of the motions of samples subjected to violent impacts by projectiles fired at them (Fig. 18.3). Interpretation recognizes that shock compression causes heating greater than adiabatic compression. It must also be noted that the compression is so sudden as to invite doubt about how nearly specimens come to hydrostatic and thermodynamic equilibrium. Anderson (1995, Chapter 12) discusses the problems and Stacey and Davis (2004) draw attention to doubt about pressure calibration. It appears that shock compression is a compromise between hydrostatic compression, described by K, and linear compression, represented by the higher modulus, χ. A plausible explanation is that stronger (faster) shocks give less time for adjustment towards a hydrostatic state, giving the impression of bulk modulus increasing faster with pressure than it really does and leading to overestimates of pressure. Ignoring these doubts, there is a reasonably simple interpretation of compression in terms of the speeds of the advance of a shock front through a specimen and of the following shocked material.

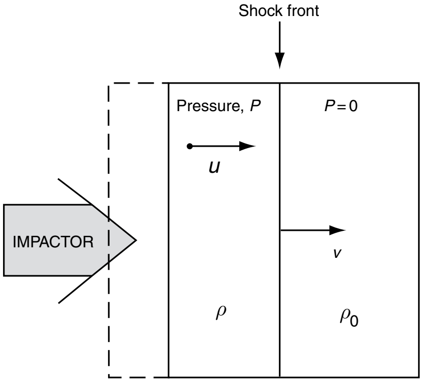


FIGURE 18.3 Geometry of shock wave propagation. A high speed impact from the left initiates a shock wave travelling at speed v through the initially stationary sample. The shock-compressed material, density ρ, moves at speed $u < v$. The unshocked material, density ρ_0, has not yet moved. The broken line indicates the extent of the sample before impact.

Applying the principle that the mass of material is conserved as the shock front advances through it, at speed v relative to the uncompressed material but speed $(v - u)$ relative to the shock-compressed material, to which speed u is imparted, we have

$$v\rho_0 = (v - u)\rho. \tag{18.5}$$

By observing both v and u by high speed photography the compression is measured:

$$\frac{\rho}{\rho_0} = \left(1 - \frac{u}{v}\right)^{-1}. \tag{18.6}$$

The same observations give the pressure generated by the shock because this is equated to the rate of change of momentum per unit area of shock front,

$$P = \rho_0 v u. \tag{18.7}$$

That is, per unit time a mass $(\rho_0 v)$ per unit area is given speed u. A series of observations with different shock intensities, produced by a range of impactor speeds, gives a $P(\rho)$ curve.

The compression is dynamic and is not, in general, adiabatic. The heating can be calculated by appealing to the conservation of energy through a shock front, across which there is a pressure difference P, causing acceleration of the material to speed u. Since the pressure is exerted by the compressed material, which advances at speed u, the power per unit area of shock front is Pu. This is equated to the rate at which energy is imparted to the material, being the sum of kinetic energy and increased internal energy,

$$Pu = \rho_0 v(u^2/2 + U^* - U_0^*). \tag{18.8}$$

U^* and U_0^* are the internal energies per unit mass in the compressed and uncompressed states, the asterisk being introduced to distinguish these quantities from the total internal energy U of an arbitrary mass, which is the convention adopted in Appendix E. With substitutions from Eqs. (18.6) and (18.7), we have

$$U_{\mathrm{H}}^* - U_0^* = \frac{1}{2}P_{\mathrm{H}}\left(\frac{1}{\rho_0} - \frac{1}{\rho}\right). \tag{18.9}$$

This is the Hugoniot equation, named after one of its original authors. A shock compression curve along which internal energy varies in the manner of this equation is referred to as a Hugoniot and the subscript H emphasizes that this relationship applies only to such compressions.

Equation (18.9) may be used to estimate either adiabatic or isothermal compressions from Hugoniot data. The calculation of an adiabat is simpler and an isotherm can be calculated from it by standard thermodynamics. From Table E.2 (Appendix E), $(\partial U/\partial V)_S = -P$, so that the variation of internal energy with adiabatic compression is

$$\mathrm{d}U_S^* = -P_S \mathrm{d}V^* \tag{18.10}$$

with subscript S to indicate adiabatic variations and continuation of the asterisk for unit mass parameters. Integrating Eq. (18.10) and combining it with Eq. (18.9), we have

$$U_S^* - U_{\mathrm{H}}^* = -\int_{V_0^*}^{V^*} P_S \mathrm{d}V^* - \frac{1}{2}P_{\mathrm{H}}(V_0^* - V^*), \tag{18.11}$$

which is the difference between internal energies along an adiabat and a Hugoniot at the same density, $1/V^*$. For constant volume changes we have, from Table E.2,

$$\left(\frac{\partial U}{\partial P}\right)_V = \frac{m}{\gamma\rho} = \frac{V}{\gamma}, \tag{18.12}$$

so that, in the reasonable approximation that the Grüneisen parameter, γ (Eq. (18.2)), is independent of temperature at constant volume,

$$U_S - U_{\mathrm{H}} = (V/\gamma)(P_S - P_{\mathrm{H}}). \tag{18.13}$$

Therefore the 'correction' of Hugoniot pressure to an adiabat is given by

$$P_S = P_{\mathrm{H}}[1 - (\gamma/2)(\rho/\rho_0 - 1)] - \gamma\rho \int_{V_0^*}^{V^*} P_S \mathrm{d}V^*. \tag{18.14}$$

The integral in Eq. (18.14) assumes knowledge of $P_S(\rho)$, which it is the purpose of the calculation to determine, so use of this equation is iterative. Calculation of an isotherm from Hugoniot data, either directly or via an adiabat, assumes knowledge of the variation of γ with ρ by an appeal to one of the theories in Sections 18.3 to 18.5.

18.3 The appeal to atomic potentials

An atomic potential function is an expression for the mutual potential energy, ϕ, of neighbouring atoms as a function of their separation, r. Every finite strain theory implies a potential function with the general form of Fig. 18.4, the essential feature of which is an asymmetry about the potential minimum. It is easier to stretch atomic bonds than to compress them. There are two related consequences: bulk modulus increases with pressure and materials normally expand when heated. A finite strain theory is concerned with the first of these effects, but it is closely linked to theories of thermal properties, especially the Grüneisen parameter (Section 19.3). There is a conceptual advantage in starting a finite strain theory with a potential function; it makes the underlying physical assumptions clear and it leads naturally to the link with thermal properties. Normal elasticity theory (infinitesimal strain theory) considers very small

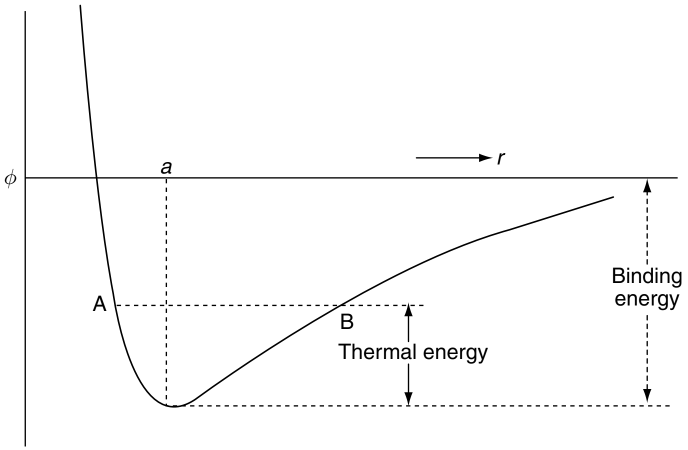


FIGURE 18.4 The form of an atomic potential function, representing the interaction energy of neighbouring atoms as a function of their separation, *r*. The equilibrium separation, *a*, is the result of a balance between attractive and repulsive forces. The increasing gradient with decreasing *r* causes the bulk modulus to increase with compression. Thermal oscillation between A and B causes thermal expansion by allowing greater bond extensions than compressions.

displacements from the minimum in Fig. 18.4. In this range ϕ varies as the square of displacement, giving forces proportional to displacement and therefore elastic moduli that can be treated as constants. There is no unique representation of ϕ for extrapolation outside this range but only numerous empirical potential functions that have been used to represent the behaviour of strongly compressed material. Computer models (molecular dynamic calculations) can handle analytically unmanageable potential functions, but their validity still depends on assumed functions There are two general forms, power law potentials and functions incorporating exponentials. They all have a minimum in φ at $r = a$, the equilibrium zero pressure spacing, at which $\mathrm{d}\phi/\mathrm{d}r = 0$. Thus density variation is given by

$$\rho/\rho_0 = V_0/V = (a/r)^3, \qquad (18.15)$$

where subscript zero indicates zero pressure values.

Consider a crystal with N atoms, each of which has $6f$ neighbours and to each of which it is bonded with energy ϕ. For a simple cubic crystal $f = 1$, for an atomic close-packed structure $f = 2$ and for the diamond structure $f = 2/3$. Since each bond is shared by two atoms, the total bond energy of the crystal is

$$E = 3Nf\phi. \qquad (18.16)$$

Let the volume of the crystal be

$$V = Ngr^3, \qquad (18.17)$$

where g is another dimensionless constant with a value of unity for a simple cubic crystal. It is the volume per atom, relative to a cube of side r; it has a minimum value of $1/\sqrt{2}$ for atomic close-packed structures and a maximum value of $8/3\sqrt{3}$ for the diamond structure. Ignoring, for the present, thermal effects, the pressure at arbitrary atomic spacing, r, is

$$P = -\frac{\mathrm{d}E}{\mathrm{d}V} = -\frac{\mathrm{d}E/\mathrm{d}r}{\mathrm{d}V/\mathrm{d}r} = -\frac{f}{g}\cdot\frac{1}{r^2}\frac{\mathrm{d}\phi}{\mathrm{d}r} \qquad (18.18)$$

and bulk modulus is

$$K = -V\frac{\mathrm{d}P}{\mathrm{d}V} = -V\frac{\mathrm{d}P/\mathrm{d}r}{\mathrm{d}V/\mathrm{d}r} = \frac{f}{3g}\left(\frac{1}{r}\frac{\mathrm{d}^2\phi}{\mathrm{d}r^2} - \frac{2}{r^2}\frac{\mathrm{d}\phi}{\mathrm{d}r}\right). \qquad (18.19)$$

We are also interested in

$$\frac{\mathrm{d}K}{\mathrm{d}P} = \frac{\mathrm{d}K/\mathrm{d}r}{\mathrm{d}P/\mathrm{d}r} = \frac{\dfrac{\mathrm{d}^3\phi}{\mathrm{d}r^3} - \dfrac{3}{r}\dfrac{\mathrm{d}^2\phi}{\mathrm{d}r^2} + \dfrac{4}{r^2}\dfrac{\mathrm{d}\phi}{\mathrm{d}r}}{3\left(\dfrac{2}{r^2}\dfrac{\mathrm{d}\phi}{\mathrm{d}r} - \dfrac{1}{r}\dfrac{\mathrm{d}^2\phi}{\mathrm{d}r^2}\right)} \qquad (18.20)$$

because we wish to find functions ϕ that match the gradients of the $K(P)$ curves for the deep Earth. For some purposes we need higher derivatives but it is generally more convenient to treat each case specifically than to pursue the general form. The algebra is simplified by noting that the denominator of Eq. (18.20) has the form of K (Eq. (18.19)), so that the next stage is obtained by differentiating the product $K(\mathrm{d}K/\mathrm{d}P)$.

The simplest potential functions are those that represent ϕ as a function of $1/r$:

$$\phi = a_1(a/r)^m + a_2(a/r)^n + a_3(a/r)^p + a_4(a/r)^q + \cdots, \qquad (18.21)$$

where a is the equilibrium spacing, a_1, a_2,... are coefficients with the dimensions of energy and n, m,... are dimensionless exponents, commonly but not necessarily integers. There are numerous variants. With just two terms and arbitrary m, n ($n > m$ so that a_1 is negative) we have the potential proposed by G. Mie before the introduction of quantum mechanics suggested different forms. If $m = 1$ to represent Coulomb attraction it is known as the Born–Mie potential. If $m = 6$, $n = 12$ we have the Lennard-Jones potential for dipole–dipole interactions. J. Bardeen proposed a three-term potential with exponents 1, 2, 3 and the theory of Birch (1952) can be represented by Eq. (18.21) with either 2, 3 or 4 terms having exponents 2, 4, 6, 8. With all four terms it is referred to as the fourth-order theory, although there cannot in principle be fewer than two terms, so, with this nomenclature, there is no first-order theory. With exponents that are all multiples of 2, it is evident that Eq. (18.21) is a polynomial in $(\rho/\rho_0)^{2/3}$ so that repeated differentiation is straightforward. Writing $(\rho/\rho_0) = x$, differentiating Eq. (18.21) with respect to x as far as $K'' = \mathrm{d}^2K/\mathrm{d}P^2$ and eliminating coefficients by substitution of zero pressure ($x = 1$) values of K, K' and K'', the fourth-order Birch theory gives

$$\begin{aligned} P &= -\,\mathrm{d}E/\mathrm{d}V = (x^2/V_0)\mathrm{d}E/\mathrm{d}x \\ &= (9/16)K_0(-Ax^{5/3} + Bx^{7/3} - Cx^3 + Dx^{11/3}), \end{aligned} \qquad (18.22)$$

where

$$\begin{aligned} A &= K_0K_0'' + (K_0' - 4)(K_0' - 5) + 59/9, \\ B &= 3K_0K_0'' + (K_0' - 4)(3K_0' - 13) + 129/9, \\ C &= 3K_0K_0'' + (K_0' - 4)(3K_0' - 11) + 105/9, \\ D &= K_0K_0'' + (K_0' - 4)(K_0' - 3) + 35/9. \end{aligned} \qquad (18.23)$$

For the third-order theory, a_4 and D are assumed to be zero so that

$$K_0K_0'' = -(K_0' - 4)(K_0' - 3) - 35/9,$$

simplifying A, B, C:

$$\begin{aligned} A &= -2(K_0' - 4) + 8/3; \\ B &= -4(K_0' - 4) + 8/3; \\ C &= -2(K_0' - 4) \\ &= B - A. \end{aligned} \qquad (18.24)$$

For the second-order theory, with a_3 and C also zero,

$$K_0' = 4; \; K_0K_0'' = -35/9; \; A = B = 8/3. \qquad (18.25)$$

Advantages and limitations of the Birch theory and reasons for its prominence in geophysics are discussed in Section 18.4.

Doubts about the fundamental validity of equations with the power law form of Eq. (18.21) arose as soon as quantum mechanics suggested that at least the repulsive term of a potential function should have an exponential form. Early theoretical attempts to develop equations of this form were based on studies of vibrational spectra of diatomic molecules, first by P. N. Morse in 1929 and then by R. Rydberg in 1932. The Rydberg potential, which now receives more attention, is

$$\phi = A[1 - f(1 - r/a)]\exp[\zeta(1 - r/a)]. \qquad (18.26)$$

Differentiation, as for the Birch theory above, gives the corresponding finite strain equation,

$$P = 3K_0x^{2/3}(1 - x^{-1/3})\exp[\zeta(1 - x^{-1/3})], \qquad (18.27)$$

where zero pressure conditions fix A and $\zeta = (3/2)(K'_0 - 1)$. This equation was given strong support by Vinet *et al.* (1987) and is sometimes referred to as the Vinet equation, although it antedated his work by many years. However, for application to the pressures in the deep Earth the Morse and Rydberg potentials share a crippling shortcoming. In Section 18.6 we refer to the thermodynamic requirement that a finite strain theory must give a value of K' that exceeds 5/3 as $P \to \infty$. This is the quantity K'_∞ which is equal to the highest exponent of x in equations such as 18.22 or 18.27. Equation (18.27) gives $K'_\infty = 2/3$ and so fails the thermodynamic criterion by a wide margin, but Stacey (2005) pointed out that a very simple modification overcomes this problem and gives sensible fits to terrestrial data. He suggested a generalized Rydberg equation with arbitrary K'_∞,

$$P = 3K_0 x^{K'_\infty}(1 - x^{-1/3})\exp[\zeta(1 - x^{-1/3})], \quad (18.28)$$

where now

$$\begin{aligned}\zeta &= (3/2)K'_0 - 3K'_\infty + 1/2\\ &= -3K_0K''_0 - (3/4)K'^2_0 + 1/12. \quad (18.29)\end{aligned}$$

The approach to finite strain via potential functions is fundamental and leads naturally to the relationship to thermal properties (Section 19.3). However, while it is useful to remember that every finite strain theory implies a potential function, the form of the required function is not known precisely enough to account adequately for properties that depend on high derivatives of it. One such property is $K' = \mathrm{d}K/\mathrm{d}P$ and especially its infinite pressure asymptote, for which a thermodynamic argument in Section 18.6 leads to a lower bound, 5/3 (Eq. (18.56)). Most theories give fixed values for this parameter, with a wide range, mostly below this limit (Stacey, 2005, Table 1). A quite different approach to finite strain, designed to avoid this difficulty, is presented in Section 18.5.

18.4 Finite strain approaches

To most geophysicists 'finite strain' means the theory of Birch (1952) that leads to Eq. (18.22), but from a quite different starting point. It originated from an attempt by Love (1927) to extend conventional elasticity theory. This defines strain as a fractional change in a dimension, $\Delta l/l_0$, which is assumed to be infinitesimal. In his extension of the theory to finite deformation, Love imposed the mathematical requirement that, when treated in three dimensions, strain should appear invariant with rotations or interchanges of coordinate axes. His method of achieving this was to define strain in terms of the squares of the separations of material points. Elastic shear strains are very small in all situations in the Earth, so we are interested in finite strain only for the case of hydrostatic compression, with extension (or compression) the same in all directions. Then, if the separation of two material points is S_0 in the unstrained state and becomes S in the strained state, the Love-defined strain, ε_L, is given by

$$(S/S_0)^2 = 1 + 2\varepsilon_L \quad (18.30)$$

and the ratio of volumes in the strained and unstrained states is

$$V/V_0 = (S/S_0)^3 = (1 + 2\varepsilon_L)^{3/2}, \quad (18.31)$$

so that

$$\varepsilon_L = [(V/V_0)^{2/3} - 1]/2. \quad (18.32)$$

When strain energy is written as a polynomial in ε_L, starting with ε_L^2, it is not convergent. Prompted by a comment by F. D. Murnaghan, Birch (1952) redefined strain relative to the strained state, that is $\Delta l/l$ instead of $\Delta l/l_0$, and found that this greatly improved the convergence. For convenience in dealing with compression, Birch also reversed the sign of strain to make it positive for compression. With these changes the Birch-defined strain is

$$\varepsilon_B = [(V_0/V)^{2/3} - 1]/2 = [(\rho/\rho_0)^{2/3} - 1]/2. \quad (18.33)$$

ε_L is referred to as Lagrangian strain and ε_B as Eulerian strain. In the limit of very small strains both converge to the conventional definition in elasticity theory (but with a sign difference). In the Birch theory strain energy is written as a polynomial in ε_B,

$$E_S = c_2{\varepsilon_B}^2 + c_3{\varepsilon_B}^3 + c_4{\varepsilon_B}^4 + \cdots \quad (18.34)$$

with the implication that this is an infinite series, but not usable beyond the ε_B^4 term. Substituting for ε_B by Eq. (18.33) and multiplying out the terms, we see that Eq. (18.34) is a polynomial in $(\rho/\rho_0)^{2/3}$ and so is completely equivalent to the 2, 4, 6, 8 power law potential that gives Eq. (18.22). Although the Birch theory is generally presented in terms of ε_B, this is algebraically much less convenient than Eq. (18.22).

The claim for convergence that made Eq. (18.34) interesting to geophysicists is based on the observation that for the second-order theory ($c_3 = 0$, $c_4 = 0$) $K'_0 = 4$ (Eq. 18.25) and this is not far from the values for many minerals.

Thus, for these minerals $c_3 \ll c_2$. However, when lower mantle data are fitted to the Birch theory they give $c_4 \approx c_2$. Equation (18.34) is not intrinsically convergent but relies on $\varepsilon_B \ll 1$. Several authors have drawn attention to difficulties with the Birch theory. For example, if Eq. (18.22) is fitted to lower mantle or core data it gives negative D or, if the third order theory ($D = 0$) is used, it gives positive C, and in either case this means negative pressure at strong compressions. But the Birch theory survives because the geophysical community has not been convinced that there is anything better. So, we draw attention to Eq. (18.28) and the theories in Section 18.5.

Another strain-based theory was presented by Poirier and Tarantola (1998), who pointed out that there is no logical reason to define strain relative to either strained or unstrained states. If the concept of strain is used it should be defined as a small increment in deformation relative to the ambient state and integrated over the total range. This gives the total strain as a logarithm of the ratio of final and initial states, which, for volume compression, is

$$\varepsilon_{\mathrm{H}} = (1/3)\ln(V/V_0). \qquad (18.35)$$

The factor 1/3 arises because strain is defined as a change in a linear dimension. The subscript H acknowledges that Poirier and Tarantola referred to ε_{H} as Hencky strain, adopting the definition from structural geology, in which it represents large inelastic deformations. Writing strain energy as a polynomial in ε_{H}, as for ε_{B} in Eq. (18.34), they obtained their logarithmic finite strain equation,

$$P = xK_0[\ln x + (1/2)(K_0' - 2)(\ln x)^2 + \cdots]. \qquad (18.36)$$

This is the third-order equation, terminating at $\varepsilon_{\mathrm{H}}^3$. The fourth-order term is given by Stacey and Davis (2004), but is doubtfully useful, justifying the claim that Eq. (18.36) is more strongly convergent than Eq. (18.22). Although, for most purposes, the logarithmic equation is better than the Birch equation, in the extreme pressure limit it gives $K_\infty' = 1$ for all orders, falling short of the thermodynamic lower bound, 5/3, discussed in Section 18.6.

18.5 Derivative equations

The equations discussed in Sections 18.3 and 18.4 have trouble with derivative properties, in particular K' and especially its infinite pressure asymptote, K_∞'. In this section we consider equations that are derived by starting with physical arguments about the behaviour of K' and integrating to obtain $P(\rho)$ relationships, instead of working the other way. Most finite strain theories consider isothermal pressure derivatives, but in geophysics adiabatic derivatives are often more directly useful. Equations in this section apply equally well to either and subscripts T and S are dropped.

We can regard Murnaghan's equation, $K' = \text{constant}$, as a special case and the precursor to a class of equations that we refer to as K-primed equations or derivative equations. With $K' \equiv \mathrm{d}K/\mathrm{d}P = K_0'$, $K = K_0 + K_0'P$, integration gives

$$\rho/\rho_0 = (K/K_0)^{1/K'} = (1 + K_0'P/K)^{1/K'}. \qquad (18.37)$$

As Fig. 18.1 shows, this is a sufficiently good approximation to be useful over moderate pressure ranges and it is easy to apply. However, closer inspection of the figure shows that the gradients of the lower mantle and core graphs decrease with pressure, that is K'' is negative and not zero as assumed by the Murnaghan equation. Since the equations of the previous two sections give negative K'', Murnaghan's equation itself is not a step forward, but is merely a pointer to a new direction.

The first real insight on the behaviour of K' appears in a paper by Keane (1954). Keane recognized that K' decreases from its zero pressure value, K_0', towards a finite limit, K_∞', as $P \to \infty$. Although he had no direct evidence of values of K_∞', he argued that it must be bounded by limits

$$(K_0' - 1) > K_\infty' > K_0'/2, \qquad (18.38)$$

which we refer to as Keane's rule. Evidence that we now have strongly supports this rule (Section 18.9). Derivation of a tighter limit is one of the current challenges in finite strain theory because if K_∞' is known, or is related to K_0', then fitting of equations such as (18.28) or those in this section requires one fewer fitting constant and is correspondingly more secure. By making the assumption that pressure is a quadratic function of Birch strain (Eq. (18.33)), with related coefficients, Keane derived the relationship that we refer to as Keane's equation:

$$K' = K'_0 + (K'_0 - K'_\infty)K_0/K, \tag{18.39}$$

which integrates to give

$$K/K_0 = 1 + (K'_0/K'_\infty)(x^{K'_\infty} - 1), \tag{18.40}$$

$$P/K_0 = (K'_0/K'^2_\infty)(x^{K'_\infty} - 1) - (K'_0/K'_\infty - 1)\ln x \tag{18.41}$$

with $x = \rho/\rho_0$. In manipulating these equations it is convenient to note the simple forms for higher derivatives,

$$KK'' = -K'(K' - K'_\infty), \tag{18.42}$$

$$K^2K''' = K'(K' - K'_\infty)(3K' - K'_\infty). \tag{18.43}$$

Although Keane's equation has been used very occasionally its merit has not been widely recognized. It is one of the equations to be taken seriously.

The next step in development of K-primed equations was recognition that

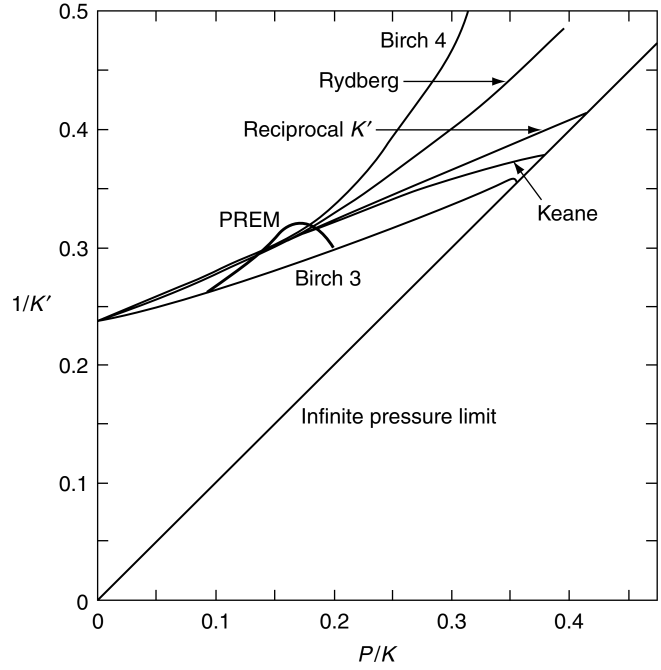
$$K'_\infty = (P/K)^{-1}_\infty, \tag{18.44}$$

which is a universal algebraic feature of all equations for which K'_∞ is positive, as proved by Stacey and Davis (2004). Although this is a standard condition of all potentially useful equations, serious use of it is possible only if a finite strain equation is written as a relationship between K' and (P/K), giving it a fixed end point at $P \to \infty$. Several such equations have been tried. The most successful is the 'reciprocal K-primed equation'

$$1/K' = 1/K'_0 + (1 - K'_\infty/K'_0)P/K. \tag{18.45}$$

By writing it in this form Eq. (18.44) is automatically incorporated. Eq. (18.45) is best represented as a graph of $1/K'$ vs P/K, as in Fig. 18.5. On this graph Eq. (18.45) is a straight line from $1/K'_0$ at $P/K = 0$ to its intersection with Eq. (18.44), which is marked 'Infinite pressure limit', a straight line of gradient 1 through the origin, along which the end points of all equations must lie. Stacey and Davis (2004) presented a method of integrating Eq. (18.45). Its integral and derivative forms are

FIGURE 18.5 Plots of $1/K'$ ($=\mathrm{d}P/\mathrm{d}K$) vs P/K for five equations fitted to the PREM model of the lower mantle (for the radius range 3630 km to 5600 km). Birch 4 (Eq. (18.23)); Birch 3 (Eq. (18.23) with three terms); Rydberg (Eq. (18.27)); Keane (Eq. (18.39)); reciprocal K-primed (Eq. (18.45)).



$$K/K_0 = (1 - K'_\infty P/K)^{-K'_0/K'_\infty}, \tag{18.46}$$

$$\begin{aligned}\ln(\rho/\rho_0) = &-(K'_0/K'^2_\infty)\ln(1 - K'_\infty P/K)\\ &-(K'_0/K'_\infty - 1)P/K,\end{aligned} \tag{18.47}$$

$$KK'' = -(K'^2/K'_0)(K' - K'_\infty), \tag{18.48}$$

$$K^2K''' = (K'^3/K'^2_0)(K' - K'_\infty)(3K' - 2K'_\infty + K'_0). \tag{18.49}$$

Although Eqs. (18.39) and (18.45) appear very different, in fact they are sufficiently similar that we can choose between them only on the basis of convenience of use. This is seen by comparing the higher derivative relationships and noting that Eqs. (18.42) and (18.48) become identical at $P = 0$, so that the relationship between K'_0, $K_0K''_0$ and K'_∞ is the same for both equations. For applications to laboratory compression data on materials that exist at $P = 0$, and for which ρ_0 and K_0 are known, use of Eq. (18.41) is generally more convenient. On the other hand, for an Earth model such as PREM, with K tabulated but ρ_0 and K_0 to be estimated by extrapolation, Eq. (18.45) has the advantage that P/K can be treated as an observed quantity and Eq. (18.47) can be fitted without involving K_0, requiring one fewer fitting constant and giving correspondingly greater certainty. K_0 is then obtained from Eq. (18.46), with K'_0 and K'_∞ already fixed by Eq. (18.47). Alternatively, K_0 can be fitted by Eq. (18.46) without involving ρ. Equation (18.45) and its integral forms were used for the data fits in Appendix F, referred to in Section 18.9.

The use of P/K as the pressure parameter in the Eqs. (18.45) to (18.47) has other important advantages. We refer to it as normalized pressure. Unlike P itself, P/K 'saturates' at a finite value with indefinitely strong compression, as in Eq. (18.44). Properties such as the thermodynamic Grüneisen parameter, that approach finite limits at extreme pressure, are much better related to P/K than to P or even ρ. At the bottom of the lower mantle the value of P/K is half of its infinite pressure limit and throughout the core it is much nearer to the infinite pressure limit than to $P/K = 0$. This means that a theoretical restriction on the infinite pressure limit, as considered in Section 18.6, is an important constraint on the very high pressure behaviour of an equation of state. The divergence of alternative equations is seen in Fig. 18.5; thermal properties calculated from higher derivatives of these equations (Section 19.3) are very different and are useful only if equations with satisfactorily constrained derivatives are applied.

We need to emphasize that all finite strain theories are empirical, whether dressed up in apparently sophisticated theoretical arguments or not. There is no unique agreed theory, but a choice must be made on the basis of plausibility arguments and convenience of use in particular applications. However, for the deep Earth the choice is limited. Only the equations for which K'_∞ is an adjustable parameter can give plausible derivative properties. New theoretical constraints are urgently needed and Section 18.6 gives an indication of what may be possible.

18.6 Thermodynamic constraints

Many of the thermodynamic identities in Appendix E connect elastic and thermal properties. They are often used in mineral physics to assess the validity of approximations and assumptions about material properties at high pressures and temperatures. A particular example that is central to the subject is the relationship between the pressure dependence of thermal expansion coefficient, α, and the temperature dependence of bulk modulus, K_T,

$$(\partial\alpha/\partial P)_T = (1/K_T{}^2)(\partial K_T/\partial T)_P = -\alpha\,\delta_T/K_T, \tag{18.50}$$

where δ_T is one of the second derivative parameters defined in Table E.1. This is typical of the conventional presentation of thermodynamic relationships in that it considers isothermal compression, both in $(\partial\alpha/\partial P)_T$ and in the use of $K_T = -V(\partial P/\partial V)_T$. For application to geophysics we are usually more interested in the variations in properties on an adiabat, and one purpose of the tables in Appendix E is to make adiabatic properties as readily accessible as the more usually quoted isothermal ones.

In the limit of extreme pressure many of these relationships simplify and the simplified forms impose constraints that equations of state must satisfy. This means that we are considering the extrapolation of equations to $P \to \infty$ and need to be careful about the implications of the extrapolation. The extrapolated values of material properties at infinite pressure are simply equation of state parameters. They are not directly observable, even in principle, but are coefficients of equations fitted to the observed pressure range. If any familiar material were to approach infinite compression it would undergo phase transitions to exotic forms with quite different equations of state. Theoretical properties and equation of state parameters of materials at extreme pressure have no relevance to the properties of lower pressure phases. The extrapolated properties must nevertheless satisfy physical laws, in particular thermodynamic relationships, even though they refer to conditions under which the material cannot exist. The laws do not break down at the (P, T) conditions at which another phase of lower free energy appears. An obvious analogy is extrapolated zero pressure properties, ρ_0, K_0, K'_0, of materials that do not survive decompression to $P = 0$. They are often quoted in high pressure mineral physics and follow the conventional relationships, but they are not the same as the properties of the zero pressure phases of the same materials. Infinite pressure extrapolation is no different. We labour this point to allay any doubt about the legitimacy of the important conclusions that are derived from the extrapolation. We are imposing constraints on parameters of equations of state, in particular K'_∞, which appears in Eqs. (18.39) to (18.49). The practical significance of this constraint is emphasized by pointing out that in the core and the lower half of the lower mantle K' is nearer to K'_∞ than it is to K'_0. This applies quite generally to derivative properties, including the Grüneisen parameter, γ, (Eq. (18.2)). Deep in the Earth they are all much closer to their infinite pressure extrapolations than to zero pressure values.

Consider the product ($\gamma\alpha T$). This product appears in many of the identities in Appendix E, including Eqs. (E.1) to (E.3), which relate adiabatic and isothermal bulk moduli and their pressure derivatives. For solids and liquids it is a small quantity, even at temperatures of the Earth's interior, and it decreases with pressure. As we demonstrate, it vanishes identically at $P \to \infty$, simplifying these identities. Equation (E.15) combines adiabatic derivatives of γ, α and T, from Table E.3, with substitutions by other identities in Table E.4 and, like all of the equations from which it is derived it is an identity, without approximation. Noting the proof by Stacey and Davis (2004) that if K' remains positive and finite then $P \to \infty$ means $V \to 0$, we can integrate Eq. (E.15) from $P = 0$ ($V = V_0$) to $P = \infty$ ($V = 0$),

$$\int_{(\gamma\alpha T)_0}^{(\gamma\alpha T)_\infty} \frac{\mathrm{d}(\gamma\alpha T)}{\gamma\alpha T(1+\gamma\alpha T)} = \ln\left[1+\frac{1}{(\gamma\alpha T)_0}\right] - \ln\left[1+\frac{1}{(\gamma\alpha T)_\infty}\right]$$

$$= \int_{V_0}^{0} (\delta_S + q)\frac{\mathrm{d}V}{V}. \qquad (18.51)$$

($\delta_S + q$) must remain finite, so the volume integral in Eq. (18.51) is $-\infty$. The only way this can be achieved is by $(\gamma\alpha T)_\infty = 0$ identically. (Stacey and Davis showed that the other apparent alternative, $(\delta_S + q)_\infty = 0$, implies the ideal gas equation, which is not relevant to solids.) Vanishing ($\gamma\alpha T$) means not only $K_{S\infty} = K_{T\infty}$ but also $K'_{S\infty} = K'_{T\infty}$. Moreover, these equalities apply at any temperature because the proof does not depend on which adiabat is considered, so all these quantities become independent of temperature. Thus K'_∞ is an unambiguous equation of state parameter, with the same value for all adiabats and all isotherms of any particular material. But it is a material constant, not a universal one. We note that Eq. (E.14) is an adiabatic derivative, so the integration that gives Eq. (18.51) follows an adiabat. As we discuss below, γ remains finite at $V \to 0$, and from Table E.2, $\gamma = -(\partial \ln T/\partial \ln V)_S$, so $T \to \infty$ on an adiabat as $V \to 0$. In spite of this $(\gamma\alpha T) \to 0$. Thus α decreases more strongly than T increases, making the product αT vanish at $V \to 0$.

Now consider the definition of q:

$$q = (\partial \ln \gamma / \partial \ln V)_T. \qquad (18.52)$$

The physics of γ is discussed in Section 19.3. Since it remains finite at $V \to 0$, by Eq. (18.52) $q_\infty = 0$. This is true however the infinite pressure condition is approached so that the corresponding adiabatic derivative, q_S, also vanishes and therefore, by Eq. (E.8), $C_S' \to 0$ at $V \to 0$. With the conditions $\delta_{S\infty} > 0$, $q_\infty = 0$, $C_{S\infty}' = 0$, Eq. (E.4) gives

$$K_\infty' > 1 + \gamma_\infty. \tag{18.53}$$

Now we can appeal to a relationship between γ and K' that is discussed in Section 19.3. An early theory by Slater (1939) made the simplifying assumption that for all elastic moduli, X, $\mathrm{d}\ln X/\mathrm{d}V$ had the same value (Poisson's ratio, ν, independent of pressure), in which case Eq. (19.32) becomes

$$\gamma_S = K'/2 - 1/6. \tag{18.54}$$

This is known as Slater's gamma and the subscript $_S$ distinguishes it from other expressions for γ. Under normal conditions ν increases with pressure and γ_S overestimates γ, but at $P \to \infty$ Slater's assumption becomes valid (Section 18.8) and therefore

$$\gamma_\infty = K_\infty'/2 - 1/6. \tag{18.55}$$

Although Eq. (18.54) is unsatisfactory, Eq. (18.55) is secure for reasons considered below. With Eq. (18.53) this means

$$K_\infty' > 5/3; \gamma_\infty > 2/3. \tag{18.56}$$

This is the thermodynamic bound on K_∞', referred to in Sections 18.3 and 18.4. The coupling to γ_∞ has proved somewhat contentious, so we examine how rigorous and general it is. Equation (18.53) is inescapable, being thermodynamically rigorous and applicable to all materials. The remaining step is an appeal to Eq. (18.55). When Eq. (18.44), which is an algebraic identity if K' remains positive, is applied to Eq. (19.39), one of the standard relationships for γ, it yields Eq. (18.55) independently of any assumption about the parameter f. Equations for γ of the acoustic type (Eq. (19.33)), derived from Grüneisen's mode definition of γ (Eq. (19.18)), also reduce to Eq. (18.55), if μ and K are related by an equation such as Eq. (18.67), and this is another way of invoking Slater's assumption, used above, that Poisson's ratio is independent of pressure. Note that these bounds on K_∞' and γ_∞ refer to these quantities as equation of state parameters of ordinary materials in the observed pressure range and not to extreme pressure states that could be reached only via dramatic phase transitions.

There is scope for further constraints of this kind. With Eq. (18.55), Eq. (E.4) gives

$$\begin{aligned}\delta_{S\infty} &= [(1/\alpha)(\partial \ln K_S/\partial T)_P]_{P\to\infty} \\ &= (K_\infty' - 5/3)/2,\end{aligned} \tag{18.57}$$

and Eq. (E.17) reduces to

$$[(1/\alpha)(\partial K_S'/\partial T)_P]_{P\to\infty} = -(K_\infty' - 5/3)(K_\infty' + 5/3)/4. \tag{18.58}$$

In principle it appears that these equations suggest the possibility of a more restrictive version of Keane's rule (Eq. (18.38)) but this remains to be proved.

18.7 Finite strain of a composite material

Section 10.4 introduces the problem of elasticity of a granular material with constituents having different elasticities. For small pressure increments the mismatch of elasticities causes grain boundary stresses, so that the applied pressure is partly supported by deviatoric stresses. The result is a composite bulk modulus higher than if all the grains were individually compressed hydrostatically. In this situation the effective modulus is approximated by K_{VRH} (Eq. (10.13)). If the deviatoric stresses relax (by grain deformation), because they are very prolonged or are too great to be supported by the material strength, then all grains are subjected to the same hydrostatic pressure and the smaller Reuss modulus, K_{R} (Eq. (10.12)) applies. We need to consider the difference between K_{R} and K_{VRH} in the Earth because seismic waves impose small stresses, and use the unrelaxed modulus, K_{VRH}, but in a homogeneous layer with a wide pressure range the variation in density with depth is described by the relaxed modulus, K_{R}. Use of the Williamson–Adams equation (Section 17.5) usually relates density variation to the seismically observed modulus and neglects the difference. We examine the significance of this to the interpretation of lower mantle properties. It has no relevance to the core, which is not granular. It should be noted that we consider here properties at pressures sufficient to ensure complete closure of pores. The pressure dependence of rock properties arising from porosity at low pressure is a separate question.

Table 18.1 gives results of calculations for a mineral mix that approximates the lower mantle. It uses 290 K laboratory observations of K and K' for silicate perovskite, subscripted pv, and magnesiowustite, subscripted mw. We see that at zero pressure K_R and K_{VRH} differ by almost 2%. Properties of the same mix at pressures corresponding to the bottom and top of the lower mantle are calculated by applying the reciprocal K-primed equation (Eqs. (18.45) to (18.47)) assuming that $K'_\infty = 2.4$ for each mineral separately. This is the best estimate of the lower mantle value. While we have no assurance that it applies to the individual minerals, the error in assuming it cannot be significant. As seen in the table, the difference between K_R and K_{VRH} decreases strongly with pressure, becoming less than 0.2% at the bottom of the mantle.

Table 18.1 Relaxed and unrelaxed bulk moduli, K_R, K_{VRH} and their pressure derivatives, for a mineral mix simulating the lower mantle, with 80% perovskite and 20% magnesiowustite by volume at $P = 0$. Each mineral is assumed to obey Eq. (18.24) with K_0 and K'_0 as listed and $K'_\infty = 2.4$. Values are calculated for a 290 K isotherm at $P = 0$ and at pressures corresponding to the top and bottom of the lower mantle

P (GPa)	0	23.83	135.75
K_{pv}	264	350.14	704.89
K_{mw}	162	251.65	605.69
$(\rho/\rho_0)_{pv}$	1	1.0811	1.3456
$(\rho/\rho_0)_{mw}$	1	1.1235	1.4781
V_{mw}/V	0.2	0.1939	0.1854
K_R (GPa)	234.47	325.44	684.12
K_V (GPa)	243.60	331.04	686.50
K_{VRH} (GPa)	239.04	328.24	685.31
$(1 - K_R/K_{VRH})$	0.019 118	0.008 530	0.001 736
K'_{pv}	3.8	3.469	2.993
K'_{mw}	4.1	3.531	2.969
K'_R	4.166	3.582	3.003
K'_V	3.899	3.499	2.992
K'_{VRH}	4.032	3.540	2.997

In Earth model data, the difference between K_R and K_{VRH} is observationally indistinguishable from a temperature gradient departing from an adiabat. As in Section 17.5, we can write

$$d\rho/dP = (\partial\rho/\partial P)_S + (\partial\rho/\partial T)_P[dT/dP - (\partial T/\partial P)_S] \tag{18.59}$$

with

$$dT/dP = (1+\xi)(\partial T/\partial P)_S = (1+\xi)\gamma T/K_S, \tag{18.60}$$

where $(1+\xi)$ is the factor by which the temperature gradient differs from the adiabat. With $\alpha = -(1/\rho)(\partial\rho/\partial T)_P$, Eq. (18.59) becomes

$$d\rho/dP = (\rho/K_S)[1 - \gamma\alpha T\xi]. \tag{18.61}$$

Values of $(\gamma\alpha T)$ over the depth of the lower mantle are 0.0507 to 0.0361, so the K differences in Table 18.1 would appear equivalent to $\xi = -0.17$ to -0.05. Over the depth of the lower mantle this integrates to a temperature deficit of 90 K relative to the adiabat. It is doubtfully significant, but the Earth model PREM is consistent with a temperature excess of about 100 K over this range, so on this basis the true excess is estimated to be 200 K.

In the interpretation of lower mantle properties the difference between K_R and K_{VRH} is insufficient to be sure that it is seen, but the difference between their pressure derivatives, K'_R and K'_{VRH}, is more significant. When a mixture of minerals with different bulk moduli is compressed the less compressible constituents become increasing volume fractions of the whole. Thus their higher bulk moduli contribute increasingly to the modulus of the composite and there is a contribution to $K' = dK/dP$ arising from the varying volume fractions, in addition to the effect of K' for each constituent. In the case of K_R we can differentiate Eq. (10.12) with respect to P, noting that $dV_1/dP = -V_1/K_1$, etc., so that

$$\begin{aligned}(K'_R+1)/K_R{}^2 =& (V_1/V)(K'_1+1)/K_1{}^2 \\ &+ (V_2/V)(K'_2+1)/K_2{}^2 + \cdots,\end{aligned} \tag{18.62}$$

giving the values of K'_R in Table 18.1. It is noticeable that K'_R is significantly higher than might be expected intuitively from the K' values for the individual minerals. In the case of K'_V we must differentiate in the same way, allowing the volume fractions to vary, because we are interested in the variation of K_V over a wide pressure range. This gives

$$K'_V = (V_1/V)K'_1 + (V_2/V)K'_2 + \cdots + K_V/K_R - 1, \tag{18.63}$$

and K'_{VRH} is the average of K'_R and K'_V.

As seen in Table 18.1, the K' bias is most noticeable at $P = 0$. Thus it is in the extrapolation of the lower mantle equation of state to zero pressure that greatest care is required in matching the equation to properties of a plausible mineral mix. If the equation of state uses $P(\rho)$ data then K_R and K'_R are required, but if seismological values of K are fitted then the appropriate value of K'_0 is K'_{VRH}. The lower mantle equation of state fit in Appendix F uses K data and the value of K'_0 (4.2) is K'_S for extrapolation on the geotherm at a temperature estimated to be 1700 K. With the temperature dependence estimated by Stacey and Davis (2004) this gives $K'_0 = 4.0$ at 290 K, in agreement with K'_{VRH} in Table 18.1.

18.8 Rigidity modulus at high pressure

Unlike compression, pure shear deformation causes no temperature rise. Adiabatic and isothermal values of rigidity modulus, μ, are identical and there are no relationships for μ corresponding to Eqs. (E.1) or (18.1). From a theoretical perspective μ is more difficult to deal with than is K and a different approach is required. But in seismology μ is as well observed as is K, so we need a fundamental understanding of it to make full use of the available data. The approach presented here is based on what is termed second-order elasticity theory. It is a conventional elasticity theory in the sense that shear strains are treated as infinitesimal (we do not have to consider finite shear strains), but a finite hydrostatic compression is superimposed. Then elastic strain energy must be calculated to second order in shear strain to derive a valid expression for μ.

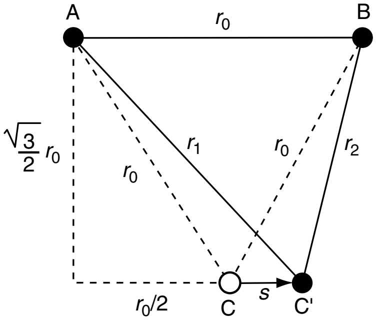


FIGURE 18.6 Changes in bond lengths between atoms that are arranged in an equilateral triangle in an unstrained crystal as a shear strain is imposed.

At low pressure, close to $r = a$ in Fig. 18.4, the potential function is parabolic, that is the departure of ϕ from the minimum value is proportional to $(r - a)^2$ and K is determined by $\phi'' = \mathrm{d}^2\phi/\mathrm{d}r^2$. But, as Eq. (18.19) shows, the complete expression involves also ϕ' and this vanishes only at $r = a$, which is the zero pressure condition by Eq. (18.18). The same principle applies to the calculation of μ, but with a reversed sign for the ϕ' term. To see how this arises, consider an atomic close-packed structure, with planes of atoms linked in arrays of equilateral triangles. The bonds connecting three such atoms are illustrated in Fig. 18.6, with the bond length changes caused by shear of the structure, displacing atom C to C′. The A–C′ and B–C′ bond lengths become

$$\begin{aligned} r_1, r_2 &= \left[\left(\frac{\sqrt{3}}{2}r_0\right)^2 + \left(\frac{r_0}{2} \pm s\right)^2\right]^{1/2} \\ &= r_0 \pm \frac{s}{2} + \frac{(3/8)s^2}{r_0} + \cdots, \end{aligned} \tag{18.64}$$

where r_0 is the equilibrium bond length at arbitrary ambient pressure and not the zero pressure length, a. It is essential to the calculation to retain the s^2 term in this equation. The energy of each bond can be written as a Taylor expansion about r_0,

$$\begin{aligned} \phi(r) - \phi(r_0) =& \phi'(r_0)(r - r_0) \\ &+ \frac{1}{2}\phi''(r_0)(r - r_0)^2 + \cdots, \end{aligned} \tag{18.65}$$

drawing attention to existence of the ϕ' term in the expression for bond energy. If $P \neq 0$, that is $r \neq$ a, $\phi'(r)$ does not vanish. The strain energy, E_S, of the bonds represented in Fig. 18.6, that is the sum of the energies of the A–C′ and B–C′ bonds less the A–C and B–C energies, is then given by substitution of r_1, r_2 for r in Eq. (18.65),

$$\begin{aligned} E_S &= \phi(r_1) + \phi(r_2) - 2\phi(r_0) \\ &= \phi'(r_0)(r_1 - r_0) + \phi'(r_0)(r_2 - r_0)^2 \\ &\quad + \frac{1}{2}\phi''(r_0)(r_1 - r_0)^2 + \frac{1}{2}\phi''(r_0)(r_2 - r_0)^2 \\ &= \left[\frac{1}{4}\phi''(r_0) + \frac{3}{4}\frac{\phi'(r_0)}{r_0}\right]s^2. \end{aligned} \tag{18.66}$$

Since the shear strain energy is $(1/2)\mu s^2$, we see that ϕ' (and therefore P) appears in the expression for μ. But this should not be surprising because it also appears in the expression for K (Eq. (18.19)).

The ϕ' term in Eq. (18.66) would be missing if the s^2 term were omitted from Eq. (18.64). This is why we refer to the analysis as second-order elasticity theory. It is this ϕ' term, and the corresponding term in Eq. (18.19), that give the pressure dependence to the ratio μ/K. The two terms in Eq. (18.66) appear with the same sign, but in Eq. (18.19) they have opposite signs. This means that the pressure terms have opposite effects for the two moduli, increasing K but decreasing μ and therefore the ratio μ/K. A calculation by Falzone and Stacey (1980) for all strain orientations and bond directions in a face-centred cubic crystal showed that this conclusion applies to all cases. It is quite general. Thus, we can write both K and μ as linear combinations of ϕ'' and ϕ', with the ϕ' term proportional to $-P$ by Eq. (18.18) and appearing with opposite signs in the two cases. Elimination of ϕ'' from these equations gives a linear relationship connecting μ, K and P, $\mu = AK + BP$. Using the zero and infinite pressure conditions to fix A and B, we have the most useful form of the μ–K–P equation,

$$\frac{\mu}{K} = \left(\frac{\mu}{K}\right)_0 - \left[\left(\frac{\mu}{K}\right)_0 - \left(\frac{\mu}{K}\right)_\infty\right] K'_\infty \frac{P}{K}. \tag{18.67}$$

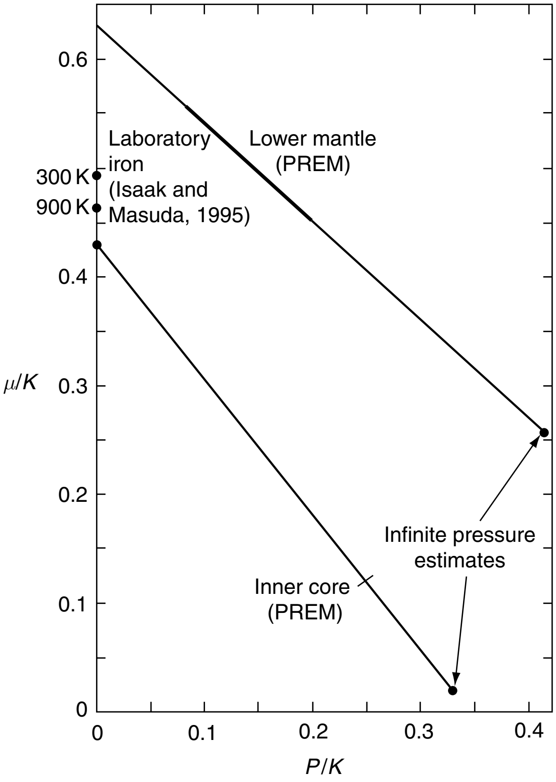


FIGURE 18.7 Lower mantle and core data fitted to the μ–K–P equation (Eq. (18.67)).

Equation (18.67) provides an explanation for the high value of Poisson's ratio, ν, that is low μ/K, of the inner core. A common inference that the inner core is close to a fluid state is quite wrong. The value of ν is just what is expected for solid iron at a pressure of 300 + GPa. But the inner core itself does not serve as a convincing test of the validity of Eq. (18.67). μ is not well observed for the inner core and the gradient of μ/K vs P/K in the PREM model is positive, not negative as Eq. (18.67) requires (Fig. 18.7). The reason for this is considered below. But a mildly surprising check on Eq. (18.67) is provided by lower mantle data, which fit the equation extremely well. With parentheses to indicate standard deviations of the final digits, the PREM tabulation gives

$$\mu/K = 0.631(1) - 0.899(6)P/K. \tag{18.68}$$

There are two interesting implications of this result. (i) Equation (18.67) is derived with the assumption that bond forces are central and it is not immediately obvious that it should apply to the lower mantle, for which non-central forces (intrinsic rigidity of angles between bonds) must be significant. (ii) The standard deviations of the coefficients in Eq. (18.68) are smaller than would be expected from the accuracy of PREM. Considering the second point first, a fit of Eq. (18.67) to the lower mantle range of the ak135 model (see Fig. 17.9) gives even smaller standard deviations but with coefficients differing from those in Eq. (18.68) by 10 standard deviations. The lower mantle is sufficiently heterogeneous for different data sets and analyses to give noticeably different coefficients, but the variations in μ and K follow Eq. (18.67) extremely well for each of them. Thus we can be confident of the validity of the form of this equation, although the coefficients are not as well determined as Eq. (18.68) suggests. So, what is the role of non-central forces, which are not recognized in the derivation, above, of Eq. (18.67)? The conclusion must be that they are not independent of the central forces but are different manifestations of the same forces.

In the case of the inner core, with very low μ/K, the central force assumption is more easily justified than in the case of the mantle, so we consider the implication of the misfit of the PREM gradient in Fig. 18.7. Although we presume μ to be diminished by anelasticity at the very high homologous temperature, T/T_M, of the inner core, this does not explain the anomalous gradient, $d\mu/dP$, which we attribute to an artifact of the Earth modelling. This gradient appears anomalously high, but a more obviously anomalous feature of PREM is the very low dK/dP in the inner core. At core pressures both solid and liquid must have close-packed atomic structures, disallowing a significant difference in dK/dP for the inner and outer cores. The well-observed (P-wave) modulus is $\chi = K + (4/3)\mu$, which is not anomalous, and the anomalies in both $d\mu/dP$ and dK/dP can be adjusted to satisfactory agreement with fundamental theory without a changing $d\chi/dP$. Both gradients can be adjusted in a compensating way without affecting the radial profile of χ or the average values of μ and K.

With this conclusion, Fig. 18.7 gives us crucial information about the equation of state for the core. We cannot extrapolate the core data past $P/K \approx 0.35$ (at which μ/K would be zero). Thus $(P/K)_\infty < 0.35$ and so, by Eq. (18.44), $K'_\infty > 1/0.35 = 2.8$. But $(P/K)_\infty$ must exceed the inner core value, ~ 0.25, so we know that $K'_\infty < 4.0$. Stacey and Davis (2004) concluded that, for the core, K'_∞ is very close to 3.0. In Section 18.9 we consider how this relates to the use of derivative equations of state, especially Eq. (18.45), for the core.

Now we draw a conclusion from Eq. (18.67) that is necessary to the infinite pressure extrapolation of the Grüneisen parameter, as in Eq. (18.55). Differentiating Eq. (18.67), we have

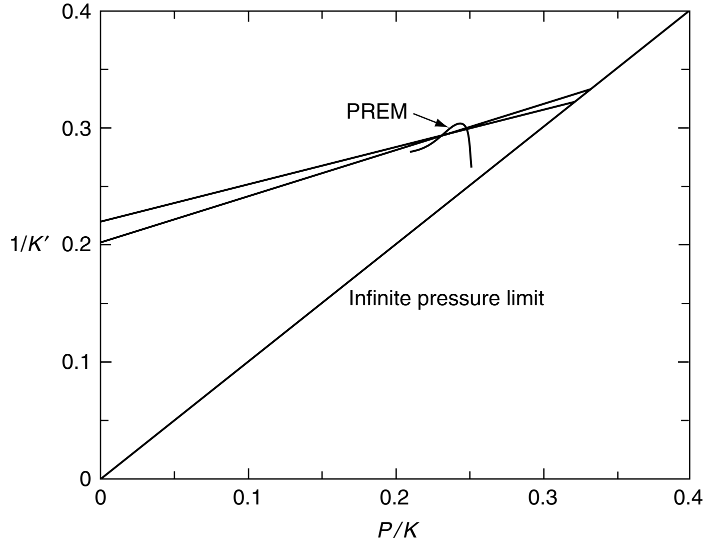
$$\frac{d\ln(\mu/K)}{d\ln P} = -\frac{1}{(\mu/K)}\left[\left(\frac{\mu}{K}\right)_0 - \left(\frac{\mu}{K}\right)_\infty\right] K'_\infty \frac{P}{K}\left(1 - K'\frac{P}{K}\right). \tag{18.69}$$

At $P \to \infty$, $(1 - K'P/K) \to 0$ by Eq. (18.44), so μ/K becomes independent of P. Since constant μ/K was assumed in Slater's derivation of Eq. (18.54), the derivation becomes valid in the $P \to \infty$ limit and, even though Eq. (18.54) is unsatisfactory, Eq. (18.55) is well founded and Eq. (18.56) follows.

18.9 A comment on application to the Earth's deep interior

Figure 18.7 draws attention to the fact that, on the P/K scale, the core is sufficiently close to the infinite pressure limit to restrict K'_∞ to a limited plausible range. The same conclusion is reached in another way from Fig. 18.8. This presents the PREM outer core data on a $1/K'$ vs P/K plot, showing its approach to the infinite pressure limit given by Eq. (18.44). Plotted in this way, the core appears much closer to the $P = \infty$ condition than to $P = 0$. The lines through the data are alternative plots of Eq. (18.45), which also show as straight lines on this figure and which intersect Eq. (18.44) at $P \to \infty$. Stacey and Davis (2004) found that these alternatives bounded the plausible range of core fits, but favoured the one with $1/K'_0 \approx 0.2$, which gives $K'_\infty = 3.0$. That Figs. 18.7 and 18.8 are both consistent with the same value of K'_∞ means that Eq. (18.45) is consistent with Eq. (18.67). There are only three finite strain equations that can accommodate the value of

FIGURE 18.8 PREM data for the outer core on a $1/K'$ vs P/K plot, showing two fits to Eq. (18.45) and the infinite pressure limit (Eq. (18.44)).



K'_∞ fixed in this way, Eqs. (18.28), (18.39) and (18.45). Although they appear very different they are in fact quite similar in form, which means that the choice between them is not very important. Equation (18.45) and its integral forms were used to calculate the lower mantle and core properties listed in Appendix F, alongside details of the PREM model that was used as the starting point for the calculations.

A particular reason for care in selecting an equation of state is that derivative properties, K' and even more so K'', differ greatly between equations and they are needed for calculation of thermal properties (Chapter 19). Realistic values of parameters, such as the Grüneisen parameter (Section 19.3), can be calculated only by using a finite strain equation for which these derivatives are satisfactory, and a glance at Fig. 18.5 shows that deep in the lower mantle the equations are very different, even though they are all fits to the PREM data. The differences are even greater in the core. However, it must be recognized that none of these equations is more than an empirical approximation. We have no rigorous theory and can only make a choice from the available equations on the basis of tests and constraints, such as those outlined in Sections 18.6 and 18.8, and convenience of use. For geophysical applications there are advantages to the use of P/K as the pressure parameter, as in Eqs. (18.45) to (18.47), rather than P/K_0, which is more conventional. One is that P and K are both tabulated in Earth models, so that P/K can be treated as an observed quantity. Another is that properties such as the Grüneisen parameter and the ratio μ/K reach finite limits as $P \to \infty$ in much the same way as does P/K and so relate more naturally to it. Stacey and Davis (2004) give a more extended discussion of this problem and used Eqs. (18.46) and (18.47) for the equation of state fits to the lower mantle and core listed in Appendix F.

Equation-of-state fits to the lower mantle require a cautionary note. They implicitly assume that the material is uniform, not only in mineral structure but in phase structure, including electronic structures of individual minerals. An electronic phase transition in iron ions at lower mantle pressures is now well recognized and it modifies physical properties, including elasticity and acoustic velocities (Lin *et al.*, 2006). A theoretical overview of the problem is presented by Sturhahn *et al.* (2005), who point out that, at the high temperatures of the lower mantle, the transition is smeared out over a wide pressure range, probably most of the lower mantle. The transition involves a realignment of the spins of the 3d electrons that are responsible for the magnetic properties of iron. At low pressure the available spins in each atom are aligned parallel, giving the atoms magnetic moments. This is referred to as the high spin state. At pressures above about 60 GPa, the spins become paired to cancel the magnetic moments in a low spin state. The effect has been studied in greatest detail in (Mg,Fe)O, but also affects perovskite. Such a transition, in which increases in density and bulk modulus are spread over a wide pressure range, is difficult to distinguish seismologically from the behaviour of a constant phase structure with modified properties, such as a high value of $\mathrm{d}K/\mathrm{d}P$. The full implications for the lower mantle have yet to be worked out; this problem is discussed also in Section 17.5.

We have emphasized that equations of state are different for different materials, or different phases of the same material. Thus the properties of iron under inner core conditions do not extrapolate to the properties of laboratory iron, which has a different structure. The same limitation applies to comparison of the outer core with laboratory measurements on liquid iron. The liquids have different structures, each being a dislocated version of the solid with which it is in equilibrium, although the change in liquid structure near to a solid–solid phase transition is smeared out over a range in pressure (and varies with temperature). This was demonstrated by Sanloup *et al.* (2000) for liquid iron close to the solid δ–γ transition. Thus we cannot use laboratory observations on liquid iron to constrain the equation of state of the outer core. The values of K_0 and K'_0 are different. But for the core we have a value for K'_∞, obtained as described in Section 18.8, and, as seen in Fig. 18.8, this is nearer to the core value of K' than is K'_0. K'_∞ is a stronger constraint on the equation of state of the core than is K'_0.

19

Thermal properties

19.1 Preamble

Convection is an underlying theme for this chapter, the following four and Chapters 12 and 13. With the probable exception of the inner core, the entire Earth is convecting and this must always have been so. Many of the topics that geophysicists study, including earthquakes, tectonics and the geomagnetic field, are consequences of convection, thermally driven in the mantle but at least partly driven by a process of compositional separation in the core. The Earth is a thermodynamic engine, generating mechanical energy in the process of transferring heat from the hot interior to the surface. The sources of heat are considered in Chapter 21 and the thermodynamic efficiency and resulting mechanical power in Chapter 22. The calculations rely on estimates of the thermal properties considered in this chapter, especially the Grüneisen parameter.

There must be a general correspondence between local high temperatures and low seismic wave speeds in the mantle and this is sought by the technique of seismic tomography (Section 17.7), but the pattern of convection is not so simple as to make this straightforward, except for the observation of high wave speeds in the cool subducting slabs. Superimposed compositional variations confuse the picture. One approach is to compare the P- and S-wave speeds and for this purpose we need to know the temperature dependences of the wave speeds. It appears that, at least in the deepest part of the mantle, temperature and composition are correlated. The effects of composition on seismic wave speeds at lower mantle pressure are not well documented; the immediate task is to produce reliable information on the temperature effect.

The strongest control on estimates of core temperatures is the observation that the inner core boundary (ICB) marks a transition from liquid to solid iron alloy. This has prompted numerous experiments as well as theories to determine the melting point of iron at very high pressures. Experimentally this is very difficult and the observations that we have rely on extrapolations to ICB pressure from somewhat lower pressures. There are also theoretical difficulties. The solid phase of iron at core pressures is not the familiar low pressure body-centred cubic (α) form, or even the face-centred (γ) form to which iron transforms between 1192 K and 1617 K (and is approximated by the structure of stainless steel) but hexagonal close-packed ε-iron. The theory of simple melting (Section 19.4) can be applied to any of these phases if they are well removed from the triple points on the melting curve that mark the solid–solid phase transitions, but the theory cannot be extrapolated through the triple points. There is also doubt about the melting point depression by light solutes in the core. We estimate the ICB temperature as 5000 K, acknowledging that it could be in error by at least 500 K.

At the high temperatures of the Earth's interior, thermal properties are approximated by classical theory. In this context 'high' means above the characteristic Debye temperatures,

θ_D, of minerals that (for insulators) mark a transition in specific heat from strong temperature dependence at low T to almost temperature independent high T behaviour. Values of θ_D for common minerals at zero pressure are listed by Anderson and Isaak (1995); almost all fall within the range 200 K–1000 K, with those important to the mantle at the upper end of this range. Except for a thin surface layer, $T/\theta_D > 1$ throughout the Earth, but that is not true for laboratory conditions. Unless measurements are made at high temperatures, laboratory data on most thermal properties can be applied to the Earth only by extrapolations involving some understanding of the quantum phenomena that become obvious at very low temperatures.

The quantum theory of specific heat began with Einstein's theory of the energy of a harmonic oscillator, meaning one of the vibrational modes of a crystal. This is also the starting point for a fundamental understanding of thermal expansion. Expansion coefficient and specific heat are closely linked and the theoretical connection is the Grüneisen parameter, γ. This is a dimensionless combination of four familiar properties, as in Eq. (18.2), repeated here:

$$\gamma = \alpha K_T/\rho C_V = \alpha K_S/\rho C_P. \qquad (19.1)$$

γ can also be defined as the ratio of thermal pressure to thermal energy.

Throughout the Earth the numerical value of γ is in the range 1.0 to 1.5. It is much more nearly constant than α or K_T. It is virtually independent of temperature at high T (at constant volume) and its modest pressure dependence is a subject of several theories that have converged to reasonable agreement. The use of γ has become central to studies of the thermal physics of the Earth because it provides a control on calculations that is otherwise unavailable. In the Earth K_S/ρ is well observed and C_P is clearly understood, but the only useful values of α are obtained from γ by Eq. (19.1). Normally γ is used as a parameter in its own right and not just as a proxy for α; thermodynamic equations can be expressed in terms of either, but whichever is used the numerical values are obtained from γ. The concept of γ originated in solid state physics but it rarely rates even a passing mention in texts on solid state or thermal physics. This is a reason for giving it close attention in Section 19.3. The principal use of γ in geophysics is in the estimation of adiabatic temperature gradients and the mechanical power of convection. It is important also in the extrapolation of deep Earth properties to zero pressure (Section 18.9) and to melting theory (Section 19.4).

Many thermodynamic identities are relationships between thermal and elastic properties (thermoelasticity), and those most useful to geophysics are listed in Appendix E. They are basic theoretical tools, the algebraic rules of the subject. We often make simplifying approximations and these require some care. By starting with the identities in Appendix E, we can judge the adequacy of these approximations. One is the assumption of constant specific heat, C_V. Since derivatives of C_V appear in many of the identities, this is obviously a convenient simplification, but we need to consider how satisfactory it is. It is applicable only to insulators; the electron contribution to the specific heat of the core does not behave in the same way as the lattice component and it accounts for about 1/3 of the total. For the mantle the assumption that $C'_S \equiv (\partial \ln C_V/\partial \ln V)_S$ is negligible is much better than the assumption $C'_T \equiv (\partial \ln C_V/\partial \ln V)_T \approx 0$. This is a useful point because in geophysics we are generally more interested in adiabatic properties than in isothermal ones. As can be seen in Table E.3, particular care is needed with derivatives of α because C'_T appears in the entries for $\partial\alpha$ at constant P or V and it is divided by the small quantity $(\gamma\alpha T)$, making it an important term. For constant S, however, this problem does not arise.

The central theme of this chapter is the calculation of thermal properties of the Earth, starting with seismological data. The essential link between elastic and thermal properties is the Grüneisen parameter (Eq. (19.1)). By using it, we are able to apply the detailed information about the Earth's interior derived from seismology to the study of its thermal behaviour, especially convection and the thermodynamic basis of tectonics.

19.2 Specific heat

It is a popular supposition that the Earth is hot inside because of the heat released by radioactivity. Thus, it may come as a surprise to realize that the total radiogenic heat release in the life of the Earth is much less than either the heat loss over this time or the present stored heat. Heat capacity plays an essential role in the discussion of the global energy budget (Chapter 21) as well as appearing as a thermodynamic parameter in the equations used to study deep Earth physics.

As mentioned in Section 19.1, the interior of the Earth is hot enough for the thermal vibrations of atoms to be treated classically to a useful approximation. Ignoring for the moment the effect of conduction electrons in the metallic core, this means that the thermal energy (of an insulator) varies linearly with temperature, T. For a solid with N atoms it is $3NkT$ (minus a constant, as in Eq. 19.13, below), where k is Boltzmann's constant. We often refer to moles and, for a substance with n atoms per molecule, the thermal energy can be written as $3nRT$ per mole, where $R = N_\mathrm{A}k$ is the gas constant and N_A is Avogadro's number. But many geological materials are compounds or mixtures for which a mole is not a convenient concept and thermal energy is better expressed in terms of mass and mean atomic weight, $\overline{m}$. Then the classical thermal energy, $3RT/\overline{m}$ per gram atom, is

$$\text{Classical}\, E_\mathrm{Thermal} = 24\,943\, T/\overline{m}\ \mathrm{J\,kg^{-1}}. \qquad (19.2)$$

Note that the mole is a unit identified with the gram, but this equation gives energy per kilogram. For the mantle we estimate $\overline{m} \approx 22$ and for the outer core $\overline{m} = 44.53$ with the composition in Table 2.5, or 50.16 for the inner core.

Specific heat, the temperature derivative of thermal energy, has a corresponding classical value if heating or cooling occurs at constant volume, V,

$$\begin{aligned}\text{Classical}\, C_V &= (\partial E_\mathrm{Thermal}/\partial T)_V \\ &= 24\,943/\overline{m}\ \mathrm{J\,K^{-1}\,kg^{-1}}. \qquad (19.3)\end{aligned}$$

C_V is the specific heat occurring most often in theoretical discussions but is not directly measurable because, except for gases, heating cannot occur at constant volume. In the normal situation of heating or cooling at constant pressure, P, there is an additional factor arising from the energy required for thermal expansion and the specific heat, C_P, is related to C_V by one of the standard identities in Appendix E,

$$C_P = (\partial E_\mathrm{Thermal}/\partial T)_P = C_V(1 + \gamma\alpha T). \qquad (19.4)$$

Equation (19.4) is an identity, valid for all materials in all situations, and is not restricted to the classical (high temperature) regime, as are Eqs. (19.2) and (19.3). In the Earth, C_P exceeds C_V by 3% to 10%. Values of C_P are listed in the thermal model in Appendix G.

Equation (19.3) is the Dulong–Petit law of specific heat, which arose from early recognition that heat is energy of atomic motion. But, even within the assumption of classical physics there is a small correction. The mean kinetic energy per atom at temperature T is ½ kT for each of the three independent directions of motion (x,y,z), and the Dulong–Petit assumption is that the average potential energy of stretched and compressed atomic bonds is equal to the average kinetic energy. This is the principle of equipartition, which would apply if the bonds were perfectly harmonic, giving sinusoidal oscillations. But atomic bonds are anharmonic, as illustrated in Fig. 18.4. There is a small anharmonic correction to C_V, discussed in Section 19.8.

When we consider the thermal properties of Earth materials at laboratory temperature, the classical assumption breaks down. It is no longer satisfactory to think in terms of thermal energies of individual vibrating atoms, but, rather, the energies of vibrational modes of a crystal lattice as a whole. The atoms do not move independently; there are standing waves with a wide range of wavelengths and corresponding frequencies. These modes are not excited to any arbitrary level of vibration, as assumed in the classical theory, but are restricted by quantum principles, so that a mode of frequency ν may have any of a series of discrete energies $(n + ½)h\nu$, where n is an integer and h is Planck's constant. This is the basis of the quantum theory of specific heat. The theory may appear to be of limited relevance to the Earth, almost all of which is at a sufficiently high temperature for $kT > h\nu$, so that the energy levels are effectively blurred into a continuum, allowing the classical approximation (Eq. (19.3)) to be used. However, the mode theory is the starting point for an understanding of the Grüneisen parameter, as in the following section.

The level to which a mode is excited thermally is determined by the average thermal energy, kT, of the other modes with which it interacts. The probability of excitation to the nth level is

$$p(n) = \exp(-nh\nu/kT)/\sum_{n=0}^{\infty}\exp(-nh\nu/kT). \tag{19.5}$$

This is the Boltzmann energy distribution, with a denominator that is a normalizing factor to make the sum of all $p(n)$ equal to unity. The state n, when it occurs, contributes energy $nh\nu$, so that its contribution to the average energy of the mode, $\overline{E}$, is $nh\nu p(n)$. Summing over all states, $n = 0$ to ∞,

$$\overline{E} = h\nu\sum_{n=0}^{\infty} n\exp(-nh\nu/kT)/\sum_{n=0}^{\infty}\exp(-nh\nu/kT). \tag{19.6}$$

Writing $x = \exp(-h\nu/kT)$, the denominator of Eq. (19.6) is recognized as a geometric progression,

$$\sum_{n=0}^{\infty} x^n = 1/(1-x), \tag{19.7}$$

and the numerator is

$$\begin{aligned} h\nu\sum_{n=0}^{\infty} nx^n &= h\nu\sum_{n=0}^{\infty} x\mathrm{d}(x^n)/\mathrm{d}x \\ &= h\nu x\mathrm{d}\left(\sum_{n=0}^{\infty} x^n\right)/\mathrm{d}x = h\nu x/(1-x)^2. \end{aligned} \tag{19.8}$$

Using Eqs. (19.7) and (19.8) in (19.6), with substitution for x,

$$\overline{E} = h\nu x/(1-x) = h\nu/[\exp(h\nu/kT) - 1]. \tag{19.9}$$

This is the Einstein model of specific heat, applicable to harmonic oscillators all of the same frequency, ν. At high temperatures, $kT \gg h\nu$, it reduces to the classical situation, $\overline{E} \approx kT$, and at low temperatures, $kT \ll h\nu$, there is negligible excitation above the ground state, ½$h\nu$. The corresponding contribution to specific heat is obtained by differentiating Eq. (19.9) with respect to T at constant V, with the assumption that ν is independent of T if V is held constant,

$$\begin{aligned} C_V &= (\partial\overline{E}/\partial T)_V \\ &= k(h\nu/kT)^2\exp(h\nu/kT)/[\exp(h\nu/kT) - 1]^2. \end{aligned} \tag{19.10}$$

This reduces to k for $h\nu/kT \ll 1$ (the classical limit). The assumption that ν is constant at constant V would be valid if a lattice mode were a perfect harmonic oscillator, but, as in the classical theory referred to above, it must be seen as an approximation when applied to an anharmonic oscillator, with a potential function having a form represented by Fig. 18.4.

Crystals have vibrational modes with a wide range of frequencies and can be represented as collections of Einstein oscillators with different values of ν. There is a corresponding wide range of temperatures, below which the modes are 'frozen out', becoming thermally inactive. Thus the transition from low to high temperature regimes is more spread than in the Einstein model with a single mode frequency. With detailed knowledge of atomic bonding, and resort to numerical methods, realistic vibrational spectra can be calculated, even for quite complicated crystal structures. Figure 19.1 gives an example. It is compared with a mathematically simple, widely used reference model by P. Debye, dating from the early twentieth century.

The Debye theory treats a material as an elastic continuum, with standing waves of all orientations and wavelengths that can be fitted into it down to a limit that gives the required total number of modes, three times the number of atoms. The number of standing waves in the frequency range ν to $(\nu + \mathrm{d}\nu)$ that can be fitted into a crystal is proportional to $\nu^2\mathrm{d}\nu$ and this is the spectrum of frequencies that the Debye model assumes, with frequency related to wavelength by an averaged acoustic velocity. Since the total number of modes in an N atom crystal is $3N$, the spectrum is cut off at the frequency, ν_D, that gives this number. Thus the Debye model

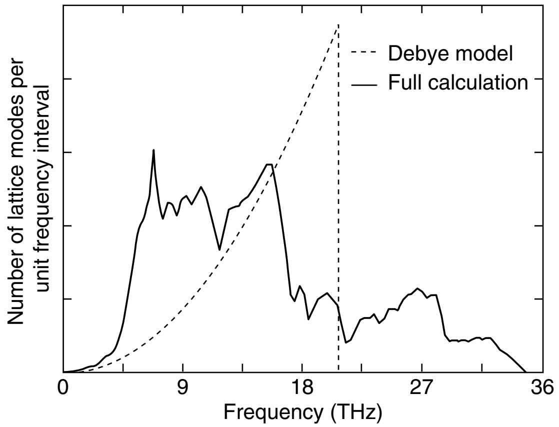


FIGURE 19.1 The spectrum of lattice modes of $MgSiO_3$ perovskite, compared with the Debye theory. Redrawn from Oganov *et al.* (2000).

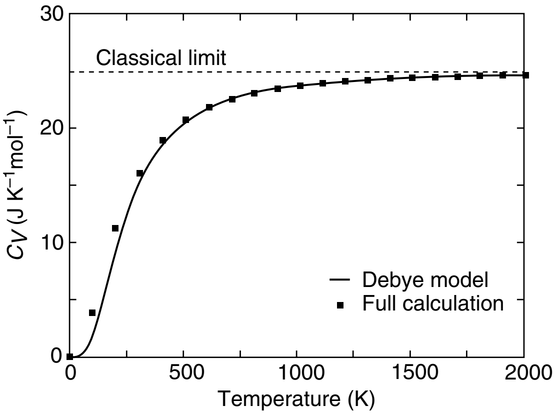


FIGURE 19.2 Specific heat of $MgSiO_3$ perovskite calculated from the lattice mode spectrum in Fig. 19.1, compared with Debye theory. Redrawn from Oganov *et al.* (2000).

has a characteristic temperature, the Debye temperature, θ_D, which is related to ν_D by

$$k\theta_D = h\nu_D. \tag{19.11}$$

At temperatures much higher than θ_D, all of the modes are fully excited and C_V approaches the classical, high temperature limit, $C_{V\infty}$, given by Eq. (19.3). At lower temperatures the high frequency modes become inactive and C_V has a reduced value. Integration of Eq. (19.10), weighted by the $\nu^2 d\nu$ spectrum, gives the Debye function, which is the sum of specific heats of a collection of Einstein oscillators (Eq. (19.10)) with a Debye spectrum of frequencies

$$\frac{C_V}{C_{V\infty}} = 3\left(\frac{T}{\theta_D}\right)^3 \int_0^{\theta_D/T} \frac{x^4 e^x}{(e^x - 1)^2} dx. \tag{19.12}$$

Although the Debye theory cannot accurately represent the fine details of thermal properties, because the spectrum of mode frequencies is not very realistic, in many situations it is a convenient approximation, especially its identification of a characteristic (Debye) temperature. The spectra in Fig. 19.1 look very different, but when both are used to calculate specific heat curves, as in Fig. 19.2, the Debye theory is seen to reproduce the essential features. The mineral represented in this figure, $MgSiO_3$, perovskite, has a high Debye temperature, about 950 K, and at laboratory temperature C_V (290 K) $\approx 0.65\, C_{V\infty}$. For all higher temperatures the discrepancy between the curves of Fig. 19.2 is quite small.

There is no analytical solution for the Debye function (Eq. (19.12)) but for $T > \theta_D$ thermal energy and specific heat can be expanded as polynomial functions of (θ_D/T):

$$E(\text{Debye}) = C_{V\infty}\left[T - \frac{3}{8}\theta_D + \frac{1}{20}\frac{\theta_D^2}{T} - \frac{1}{1680}\frac{\theta_D^4}{T^3} + \cdots\right], \tag{19.13}$$

$$C_V(\text{Debye}) = C_{V_\infty}\left[1 - \frac{1}{20}\left(\frac{\theta_D}{T}\right)^2 + \frac{1}{560}\left(\frac{\theta_D}{T}\right)^4 + \cdots\right]. \tag{19.14}$$

In estimating the total stored energy in the Earth (Chapter 21), the $(3/8)\theta_D$ term in Eq. (19.13) is important, especially as θ_D increases systematically with pressure. This term is represented by the area between the Debye curve and the classical limit extrapolated to $T = 0$ in Fig. 19.2.

It is useful to keep in mind the accuracy of the approximation that C_V is constant at high temperatures. It is not valid for metals, as discussed below, but for insulators we can differentiate Eq. (19.14) for an adequate indication:

$$\begin{aligned}(\partial \ln C_V/\partial \ln T)_V = {} & (1/10)(\theta_D/T)^2 \\ & - (3/1400)\,(\theta_D/T)^4 + \cdots .\end{aligned} \tag{19.15}$$

In the deep Earth $(T/\theta_D) \approx 1.6$, giving $(\partial \ln C_V/\partial \ln T)_V \approx 0.04$. This is no more than various uncertainties, including the mean atomic weight and the effect of anharmonicity (Section 19.8). For the Debye model $(\partial \ln C_V/\partial \ln V)_S = 0$, as demonstrated following Eq. (19.29), so by Eq. (E.6) in Appendix E $(\partial \ln C_V/\partial \ln V)_T \approx 0.05$.

Up to this point the discussion concerns only the lattice heat capacity. This is the total for insulators, including the crust and mantle, but for metals there is an additional contribution by conduction electrons. In the core it accounts for about 30% of the total specific heat. In solids the electron energy levels are spread by interactions into bands with wide energy ranges. These bands are occupied by electrons up to a level, known as the Fermi energy, set by the number of electrons. The characteristic of metals is that they have electron energy bands that overlap and are partly filled, so that electrons with energies close to the Fermi level may change states in response to an electric field (giving electrical conduction) or to thermal agitation. Some are thermally excited to energies of order kT above the Fermi level, leaving vacant states below that level. The number of electrons so excited and their average excitation energy are each proportional to T, so that the electron thermal energy is proportional to T^2 and electron specific heat is proportional to T. This is represented by

$$C_e = \beta T. \tag{19.16}$$

For laboratory iron, $\beta = 4.98\,\mathrm{mJ\ K^{-2}\,mol^{-1}} = 0.0892\,\mathrm{J\,K^{-2}\,kg^{-1}}$. This means that a temperature of 5000 K would make C_e equal to the lattice heat capacity by Eq. (19.3), but compression increases the spread of energy bands and so decreases the density of energy states at the Fermi level, reducing the value of β. Boness *et al.* (1986) reported a numerical calculation of the band structure of iron with different crystal structures over the core pressure range and summarized their results in a simple analytical approximation for β as a function of density, ρ, relative to the zero pressure value, ρ_0.

With substitution of the atomic weight of pure iron, this is

$$\begin{aligned}\beta &= (6.113\,\rho_0/\rho - 1.144)\,\mathrm{mJ\,K^{-2}\,mol^{-1}}\\ &= (0.1094\,\rho_0/\rho - 0.0205)\,\mathrm{J\,K^{-2}\,kg^{-1}}.\end{aligned} \tag{19.17}$$

The band structure and the value of β hardly depend on the crystal form and at core pressures we assume that Eq. (19.17) is a good approximation also for liquid iron. Minor amounts of lighter solute would have only a minor effect on band structure. This equation is used to estimate the electron contribution to core heat capacity in the thermal model in Appendix G, noting that Eqs. (19.16) and (19.17) give the electron contribution to C_V and that Eq. (19.4) must be used to obtain C_P. It is important also to electrical conductivity (Section 24.4) and consequently also to the thermal conductivity of the core (Section 19.6).

19.3 Thermal expansion and the Grüneisen parameter

A particular reason for geophysical interest in thermal expansion is that the dilation of heated material makes it buoyant relative to cooler, surrounding material. In an extended medium this causes convection, with an upward transfer of heat by rising hot material. Thermal convection generates the mechanical energy driving plate tectonics and the geological activity that results from it. A uniform medium in which the temperature gradient exceeds the adiabatic one, that is the rate at which temperature increases with depth due to compression, is convectively unstable. Thus convection tends to reduce any steeper gradient to the adiabatic one and through much of the Earth the temperature gradient is believed to be close to adiabatic. Calculations of this gradient (Section 19.5) and of the mechanical energy of convection (Chapter 22) rely on knowledge of the volume coefficient of thermal expansion, α, but there is no direct observation of α for the deep Earth. It is estimated indirectly from the Grüneisen parameter, γ.

The physical reason why materials expand when heated is seen in Fig. 18.4. A bond between neighbouring atoms that oscillates in length (between A and B) is extended more than it is compressed and it spends more time in the extended state because the restoring force (represented by the gradient of the curve) is then weaker. The time-averaged extension is the thermal expansion, which is a consequence of the bond asymmetry. This has another effect. As a material is compressed, the externally applied pressure can be represented in Fig. 18.4 by an added positive gradient, pushing the energy minimum to smaller r and sharpening it. This means increasing the elastic moduli. The increase in elastic moduli with pressure is well observed by seismology, which provides a direct measure of the bond asymmetry. Thus thermal expansion and the pressure dependences of elastic moduli have a common cause and can be related by an appropriate theory. The Grüneisen parameter, γ, makes this theoretical connection, which is the reason for the effort by geophysicists to understand it.

The parameter originally established in the solid state physics literature by E. Grüneisen, and bearing his name, was derived from the equations for an Einstein oscillator (Eqs. (19.9) and (19.10)). For a vibrational mode of a crystal lattice, identified as an oscillator of frequency ν_i, the Grüneisen definition is

$$\gamma_i = -(\partial \ln \nu_i / \partial \ln V)_T. \tag{19.18}$$

γ_i is the mode Grüneisen parameter, or mode gamma, and is not in general identical to the γ_i of other modes. The interest in it arises from the fact that, as shown below, an appropriate average of all of the γ_i is equivalent to the dimensionless combination of thermodynamic parameters in Eq. (19.1). In common usage, mention of the Grüneisen parameter means the definition in Eq. (19.1), but if this needs to be distinguished from Grüneisen's definition it is referred to as the thermodynamic gamma. Its usefulness in thermodynamic relationships is particularly obvious in two frequently used identities (see Table E.2 in Appendix E):

$$\begin{aligned}(\partial \ln T/\partial \ln \rho)_S &= -(\partial \ln T/\partial \ln V)_S \\ &= K_S(\partial \ln T/\partial P)_S = \gamma,\end{aligned} \tag{19.19}$$

$$(\partial P/\partial T)_V = \alpha K_T = \gamma \rho C_V. \tag{19.20}$$

Equation (19.19) gives the adiabatic temperature variation in terms of either density or pressure and Eq. (19.20) is the differential form of the Mie–Grüneisen equation, relating (thermal) pressure to thermal energy of a material heated at constant volume,

$$P_{\mathrm{Thermal}} = \rho \int_0^T \gamma C_V \mathrm{d}T \approx \gamma E_{\mathrm{Thermal}}/V, \tag{19.21}$$

where the approximate equality invokes the assumption that γ is independent of T at constant V. Over large depth ranges it varies rather little and the assumption of constant γ may be a useful approximation. Then Eq. (19.19) integrates to give

$$T_1/T_2 \approx (V_2/V_1)^{\gamma}. \tag{19.22}$$

This equation is familiar in ideal gas physics, for which $\gamma = (C_P/C_V - 1)$, because $\alpha T = 1$, simplifying Eq. (19.4). It illustrates the point that by using γ we introduce some of the simple insights of ideal gas physics to solids.

To relate γ by Eq. (19.1) to Grüneisen's definition (Eq. (19.18)), we differentiate Eq. (19.9) with respect to V at constant T and substitute for the resulting two terms by Eqs. (19.9) and (19.10), with introduction of subscript i to indicate that we are considering a single mode. This gives the volume dependence of mean mode energy,

$$\begin{aligned}(\partial E_i/\partial V)_T &= \left(\frac{\partial \ln \nu_i}{\partial \ln V}\right)_T \left\{\frac{1}{V}\frac{h\nu_i}{\exp(h\nu_i/kT)-1} - \frac{kT}{V}\frac{(h\nu_i/kT)^2\exp(h\nu_i/kT)}{[\exp(h\nu_i/kT)-1]^2}\right\} \\ &= \left(\frac{\partial \ln \nu_i}{\partial \ln V}\right)_T \left[\frac{E_i}{V} - \frac{T}{V}\left(\frac{\partial E_i}{\partial T}\right)_V\right] \\ &= -\left(\frac{\partial \ln \nu_i}{\partial \ln V}\right)_T \frac{T^2}{V}\left[\frac{\partial}{\partial T}\left(\frac{E_i}{T}\right)\right]_V.\end{aligned} \tag{19.23}$$

Now E_i is the internal energy attributable to mode i and we make use of a general identity for internal energy, U (see entries for ∂U_T and ∂P_V in Table E.2 of Appendix E),

$$(\partial U/\partial V)_T = (\partial P/\partial T)_V T - P = T^2[\partial(P/T)/\partial T]_V. \tag{19.24}$$

Identifying this expression with $(\partial E_i/\partial T)_V$ in Eq. (19.23) and dividing both sides by T^2,

$$[\partial(P_i/T)/\partial T]_V = -(\partial \ln \nu_i/\partial \ln V)_T(1/V)[\partial(E_i/T)/\partial T]_V. \tag{19.25}$$

By making the assumption that ν_i and therefore $\gamma_i = -(\partial\ln\nu_i/\partial\ln V)_T$ are independent of T at constant V, as in the Einstein theory (the starting point of the Grüneisen theory), we can integrate Eq. (19.25) with respect to T at constant V to give the relationship between thermal pressure due to mode i and its thermal energy:

$$P_i = -(\partial \ln \nu_i/\partial \ln V)_T(E_i/V). \tag{19.26}$$

This has the form of the Mie–Grüneisen equation (Eq. (19.21)), requiring Grüneisen's definition of γ_i by Eq. (19.18). Note that the Einstein assumption, that ν_i is independent of T at constant V, neglects anharmonicity (Section 19.8).

The value of γ for materials having modes with different γ_i is an average over all modes, weighted according to the heat capacities of the modes. Then, noting that γ is the coefficient relating thermal pressure to thermal energy, as in Eq. (19.12),

$$C_V\gamma = \sum C_i\gamma_i, \tag{19.27}$$

where the C_i are the various mode C_V by Eq. (19.10). If we restrict interest to high temperatures then $C_i \approx k$ (Boltzmann's constant) for each mode and γ is then a simple average of the γ_i. A temperature-dependence arises at low temperatures (additional to the small anharmonic effect of vibration amplitude on the ν_i) when modes of different γ_i contribute in different proportions because of their reduced C_i by Eq. (19.10). However, even in this situation the effect is commonly small. This is evident from a thermodynamic identity (see Tables E.3 and E.1 in Appendix E),

$$(\partial\gamma/\partial \ln T)_V = (\partial \ln C_V/\partial \ln V)_S, \tag{19.28}$$

which vanishes identically in the Debye approximation. This situation is represented by a Debye version of Grüneisen's definition of γ (Eq. (19.18)),

$$\gamma_D = -(\partial \ln \nu_D/\partial \ln V)_T = -(\partial \ln \theta_D/\partial \ln V)_T, \tag{19.29}$$

because ν_D and θ_D are related by Eq. (19.11). With the assumption that, as for all ν_i in Gruneisen's theory, ν_D is independent of T at constant V, the isothermal derivative in Eq. (19.29) is the same as an adiabatic derivative, so that in comparing Eqs. (19.19) and (19.29) we see that (T/θ_D) and therefore C_V are constant on an adiabat. Of course Debye theory is only an approximation, but it is useful in understanding why C_V is approximately constant on an adiabat.

Equation (19.18) is the starting point for a derivation of the most widely used formula for estimating γ for the Earth's interior. This is the 'acoustic gamma', γ_A, which is based on the assumption that there are only two kinds of lattice mode, corresponding to compressional and shear waves, with two shear modes for each compressional mode. This is better than the Debye method of averaging wave speeds. It assumes an isotropic medium but makes the best possible use of the seismologically observed P- and S-wave speeds, being a weighted sum of gammas for compressional (P) and shear (S) modes:

$$\gamma_A = (1/3)\gamma_P + (2/3)\gamma_S. \tag{19.30}$$

Taking mode i, controlled by elastic modulus X_i, to be a standing wave of speed $V_i = (X_i/\rho)^{1/2}$, with a wavelength λ_i that is a fixed number of lattice spacings and therefore proportional to $V^{1/3}$, the mode frequency is

$$\nu_i = V_i/\lambda_i \propto X_i^{1/2}V^{1/6}, \tag{19.31}$$

and so, by differentiation,

$$\gamma_i = -\left(\frac{\partial \ln \nu_i}{\partial \ln V}\right)_T = -\frac{1}{2}\frac{V}{X_i}\left(\frac{\partial X_i}{\partial V}\right)_T - \frac{1}{6}$$
$$= \frac{1}{2}\frac{K_T}{X_i}\left(\frac{\partial X_i}{\partial P}\right)_T - \frac{1}{6}, \tag{19.32}$$

where $K_T = -V(\partial P/\partial V)_T$ is the isothermal bulk modulus. For compressional waves $X_i = (K_S + (4/3)\mu)$ and for shear waves $X_i = \mu$, so that by Eq. (19.30)

$$\gamma_A = \frac{1}{6}\frac{K_T}{K_S + \frac{4}{3}\mu}\left[\left(\frac{\partial K_S}{\partial P}\right)_T + \frac{4}{3}\left(\frac{\partial \mu}{\partial P}\right)_T\right] + \frac{1}{3}\frac{K_T}{\mu}\left(\frac{\partial \mu}{\partial P}\right)_T - \frac{1}{6}. \tag{19.33}$$

The temperature variation through most of the Earth is believed to be close to adiabatic, so that, to the extent that seismological models give reliable gradients of the elastic moduli, they give $(\partial K_S/\partial P)_S$ and $(\partial \mu/\partial P)_S$. To convert these to the isothermal derivatives required by Eq. (19.33) we need the temperature derivatives, which are written in dimensionless forms, as defined in Appendix E,

$$\delta_S = -(1/\alpha K_S)(\partial K_S/\partial T)_P, \tag{19.34}$$

$$\varepsilon = -(1/\alpha\mu)(\partial \mu/\partial T)_P. \tag{19.35}$$

By writing a derivative identity for any parameter X,

$$(\partial X/\partial P)_T = (\partial X/\partial P)_S - (\partial X/\partial T)_P(\partial T/\partial P)_S, \tag{19.36}$$

with $(\partial T/\partial P)_S = \gamma T/K_S$ from entries in Table E.2 of Appendix E, taking X to be either K_S or μ, and using Eqs. (19.34) and (19.35) to substitute for the temperature derivatives, we obtain the required relationships:

$$(\partial K_S/\partial P)_T = (\partial K_S/\partial P)_S + \gamma\alpha T\delta_S, \tag{19.37}$$

$$(\partial \mu/\partial P)_T = (\partial \mu/\partial P)_S + \gamma\alpha T\varepsilon\mu/K_S. \tag{19.38}$$

These expressions may be obtained from entries in Table E.3 of Appendix E.

The calculation of δ_S and ε (Eqs. (19.34) and (19.35)) is a subject of Section 19.7, which is concerned with the interpretation of deep seismic velocity anomalies. For the lower mantle, for which there are sufficient data for detailed calculations, numerical values are given in Appendix G. The temperature terms in Eqs. (19.37) and (19.38) add 0.1 to 0.3 to the estimate of γ_A at deep Earth temperatures and pressures, although they have often been neglected.

A critical assessment of the assumptions and approximations implicit in the derivation of Eq. (19.33) is given by Stacey and Davis (2004), who conclude that it survives the doubts very well, whereas alternative formulae face unresolved difficulties. However, care is required in applying seismological estimates of dK/dP and dμ/dP for the deep Earth. Models such as PREM (Appendix F) represent seismic velocities and density as polynomials in radius over each of several radius ranges. This is a mathematically convenient but unphysical representation of material properties, so that differentiation to obtain dK/dP and dμ/dP yields satisfactory averages but implausible depth variations. The difficulty is overcome by fitting the PREM model to a finite strain theory (Chapter 18) for which derivative properties have been properly constrained by thermodynamic principles.

The Grüneisen theory and derivation of the acoustic formula (Eq. (19.33)) are specific to solid insulators, but the thermodynamic definition of γ (Eq. (19.1)) is applicable to every material. We need to consider its application to metals and to liquids because γ is as important to the core as it is to the mantle. There are other approaches that we now consider, but they require calibration by the acoustic calculation.

The 'free' conduction electrons in metals respond to heat as well as electric fields. Their contribution to specific heat is discussed in Section 19.2 and thermal conductivity in Section 19.6. In the core the electron heat capacity is approximately 30% of the total and there is a corresponding electron component of thermal pressure. This is related to thermal energy by an electronic Grüneisen parameter, γ_e. The total γ for a metal is the sum of lattice and electron contributions, weighted according to the contributions to specific heat in the same way as the lattice contributions are added in Eq. (19.27). In principle it is possible to calculate γ_e with a sufficiently accurate model of electron band structure, but then we do not have an independent estimate of the lattice γ because the conduction electrons contribute to elasticity. In any case most of the core is liquid, precluding use of Eq. (19.33), in which μ is prominent, so a different approach is required.

An alternative class of formulae, equally applicable to solids and liquids, has the general form

$$\gamma = [K'/2 - 1/6 - (f/3)(1 - P/3K)]/[1 - (2f/3)P/K], \tag{19.39}$$

with different values of f according to what is assumed about the thermal motions of atoms (or electrons). We refer to these formulae as the free volume type because Eq. (19.39) with $f=2$ was derived by Vashchenko and Zubarev (1963) from free volume theory. A linear model by Dugdale and MacDonald (1953), giving $f=1$, is still sometimes used. A special case, Slater's gamma, for which $f=0$ (Eq. (18.54)), is used in Section 18.6 for the infinite pressure extrapolation. Slater's formula is derived directly from Grüneisen's definition (Eq. (19.18)) with the assumption that all γ_i are equal because all mode frequencies have the same volume dependence. As in Eq. (19.32), this means that all moduli have the same volume dependence, that is Poisson's ratio, ν (not to be confused with mode frequency), is independent of pressure. Although ν is observed to increase with compression, for reasons discussed in Section 18.8, causing Slater's formula to overestimate γ, the final paragraph of Section 18.8 points out that in the infinite pressure extrapolation Slater's assumption becomes valid, making Eq. (18.55) rigorous. It provides an important constraint on finite strain theories, as outlined in Section 18.6. A difficulty with Eq. (19.39) is that we have no satisfactory theoretical value of f. An empirical solution to this problem, adopted by Stacey and Davis (2004), is to match Eq. (19.39) at $P=0$ to the zero pressure extrapolation of the acoustic formula (Eq. (19.33)). Fitted to lower mantle data, it gives $f=1.44$. A justification for this is that it gives values of γ indistinguishable from those obtained by a third theory, based on thermodynamic requirements on derivatives of γ, as discussed by Stacey and Davis (2004).

It is interesting to note that γ appears to be little affected by melting. This can be understood by recognizing thermal motion as vibrations of lattice modes that are dominated by very high frequencies (10^{12} to 10^{13} Hz). As we discuss in the following section, the fluidity of a liquid is due to the mobility of the crystal dislocations with which it is saturated, but dislocations cannot respond to stress cycles with these frequencies. So, in the consideration of lattice vibrations, the elasticities of liquid and solid are not very different.

To use Eq. (19.39) for the core we need to justify its application to metals with strong contributions to γ by conduction electrons. The point is that this equation is calculated by applying the Mie–Grüneisen equation (Eq. (19.21)), which is a model-independent identity if γ is independent of T. It simply depends on a calculation of thermal pressure and its relationship to the volume dependence of bulk modulus, and in this respect electron pressure is no different from lattice pressure. While the electron and lattice gammas may be very different, they contribute to the total in the manner of Eq. (19.27). The essential feature of Eq. (19.39) is that it makes no appeal to μ but only to K and its pressure dependence.

19.4 Melting

The standard thermodynamic identity for the variation of melting point, T_M, with pressure, P, is the Clausius–Clapeyron equation,

$$\mathrm{d}T_\mathrm{M}/\mathrm{d}P = \Delta V/\Delta S, \tag{19.40}$$

where ΔV and ΔS are the volume and entropy increments of the melting process. The derivation of this equation relies on the fact that for a substance in thermodynamic equilibrium at fixed pressure the Gibbs free energy, G (defined in Table E.1 of Appendix E), is a minimum. (At constant V the Helmholtz free energy, F, is a minimum and at constant S it is the enthalpy, H, that is a minimum.) The coexistence of solid and liquid at the melting point means that G has the same value for both states at all points on the melting curve. By considering an increment dG in G due to simultaneous increments in P and T and substituting for derivatives by entries in Table E.2 (Appendix E), we have

$$\begin{aligned}\mathrm{d}G &= (\partial G/\partial P)_T\,\mathrm{d}P + (\partial G/\partial T)_P\mathrm{d}T\\ &= V\mathrm{d}P - S\mathrm{d}T.\end{aligned} \tag{19.41}$$

Identifying T with T_M, so that the increments follow the melting curve, we require that dG be the same for solid (subscript S) and liquid (subscript L) and therefore

$$V_L \mathrm{d}P - S_L \mathrm{d}T = V_S \mathrm{d}P - S_S \mathrm{d}T, \qquad (19.42)$$

which rearranges to the form of Eq. (19.40). This is an identity, applicable to all phase transitions, and is referred to also in Section 22.3, in connection with convection through solid–solid phase transitions in the mantle.

In the case of melting, $\Delta S = S_L - S_S$ is always positive because latent heat $L = \Delta S\ T_M$ must be applied to cause melting. In what we refer to here as 'normal' or 'simple' melting ΔV is also positive, that is the liquid is less dense than the solid, so that T_M increases with P. The best known exception is water for which ΔV is negative (at low pressure), and it is useful to keep this case in mind in considering 'normal' or 'simple' melting and why there are exceptions. A successful general theory of melting is that it is a free proliferation of crystal dislocations (see Fig. 14.4) and that it occurs when the free energies of the undislocated crystal and one saturated with dislocations (identified with the liquid) are equal. Accepting this as a theory of 'simple' melting we can see why water does not fit in. The theory assumes that there is no major change in atomic coordination between the solid and liquid states, because it is not changed much by the introduction of dislocations, but common ice is structurally quite different from liquid water. As in water, strongly oriented polar bonds occur in silicates, which also do not fit well with the concept of 'simple' melting, but most metals (bismuth excepted) do so. When very high pressure is applied both solid and liquid structures become more close-packed and, regardless of low pressure behaviour, assume greater structural similarity, allowing simple melting theory to be applied with increasing confidence. The principal geophysical application is to the melting point of iron and solidification of the inner core.

Equation (19.40) is basic to the theory of melting, but it can be extrapolated to high pressures only with assumptions about both ΔV and ΔS and this is not directly useful, so that several alternative theories of melting have arisen. The one with the strongest influence on modern ideas originated with F. A. Lindemann, who deduced that melting point varied with volume and Debye temperature as $V^{2/3}\theta_D^2$. Using Eq. (19.29) and assuming $\gamma = \gamma_D$ (Eq. (19.29)), this differentiates to give (Problem 19.6)

$$(1/T_M)\mathrm{d}T_M/\mathrm{d}P = 2(\gamma - 1/3)/K_T. \qquad (19.43)$$

Equation (19.43) is often attributed to Gilvarry (1956), although comparison with his equations (38) and (31) shows that his theory gives an additional factor $[1 + 2(\gamma - 1/3)\alpha T_M]^{-1}$ on the right-hand side. The basis of its derivation was an assumption that melting occurs when the amplitude of atomic vibration reaches a critical fraction of the atomic spacing. There are now derivations of the same or very similar formulae with stronger fundamental foundations, one of which has a thermodynamic basis and is presented here. But there is an underlying, implicit assumption that the melting process is at least very similar to a proliferation of dislocations, and a simple model by Stacey and Irvine (1977) shows why. Atoms in a dislocation are displaced from their equilibrium positions so that some bonds are stretched and others compressed, but the forces are balanced. The Stacey and Irvine calculation simulated this situation with two lines of atoms locked together at their ends but having unequal numbers of atoms, so that one line was compressed and the other stretched. Bond asymmetry, as in Fig. 18.4, causes an average extension, which is identified with ΔV in Eq. (19.40), and the energy increment, relative to the undislocated lines, is identified with latent heat, $T_M \Delta S$. This allows the equation to be written in terms of derivatives of the atomic potential function and when K, P and K' are substituted by Eqs. (18.18) to (18.20), the equation has the form of Eq. (19.43) with the one-dimensional (Dugdale–MacDonald) formula for γ (Eq. (19.39) with $f = 1$). This general conclusion is fundamental and important because it shows that an equation with the form of Eq. (19.43) is independent of assumptions about dislocation structure.

We present here a more rigorous, thermodynamic derivation of the Gilvarry-type melting law. Adopting an argument by Stacey *et al.* (1989), if we consider a mass m of solid to be heated at constant volume (without melting) then the applied heat $mC_V\Delta T$ is (neglecting anharmonicity) equipartitioned between increases in atomic kinetic energy ΔE_K and bond potential energy ΔE_P, so that

$$\Delta E_K = (1/2)mC_V\Delta T. \qquad (19.44)$$

When the same temperature increment is applied at constant pressure, heat $mC_P\Delta T$ is required, but since the temperature increment is the same, ΔE_K is still given by Eq. (19.44) and therefore

$$\Delta E_P = m(C_P - C_V/2)\Delta T. \qquad (19.45)$$

With the relationship between C_P and C_V (Eq. (19.4)) this becomes

$$\Delta E_P = (1/2)mC_V\Delta T(1 + 2\gamma\alpha T). \qquad (19.46)$$

This is the increment in average bond potential energy caused by heating at constant P. At the same time there is thermal expansion,

$$\Delta V = \alpha V\Delta T, \qquad (19.47)$$

and the ratio is

$$\begin{aligned}\Delta V/\Delta E_P &= 2\alpha/(m/V)C_V(1 + 2\gamma\alpha T)]\\ &= 2\alpha/[\rho C_V(1 + 2\gamma\alpha T)],\end{aligned} \qquad (19.48)$$

which, with the definition of γ (Eq. (19.1)), is

$$\Delta V/\Delta E_P = 2\gamma/[K_T(1 + 2\gamma\alpha T)]. \qquad (19.49)$$

Now melting (at constant pressure) is an application of heat that causes no temperature rise so that the latent heat is entirely bond potential energy and Eq. (19.49) can be identified with Eq. (19.40) by writing $\Delta E_P = T_M\Delta S$, so that

$$(1/T_M)\mathrm{d}T_M/\mathrm{d}P = 2\gamma/[K_T(1 + 2\gamma\alpha T_M)]. \qquad (19.50)$$

This is similar to Eq. (19.43), which, as mentioned in the discussion of this equation, is often attributed to Gilvarry (1956) but does not correctly quote him. His theory leads to an equation with the form of Eq. (19.50) but with (γ −1/3) replacing γ.

Equation (19.50) can be rewritten to represent the variation of T_M with density at melting by applying a melting curve modulus, derived by Stacey *et al.* (1989),

$$K_M = \rho(\partial P/\partial\rho)_{T=T_M} = K_T(1 + \alpha K_M\mathrm{d}T_M/\mathrm{d}P). \qquad (19.51)$$

Combined with Eq. (19.50) it yields a simpler result:

$$\mathrm{d}\ln T_M/\mathrm{d}\ln\rho = 2\gamma. \qquad (19.52)$$

Both Eqs. (19.50) and (19.52) slightly overestimate the pressure dependence of T_M. They rely on the assumption that bonds are stretched and compressed but not broken. In the melting of atomic close-packed structures roughly 2% of atomic bonds are broken and this is a measure of the errors in these equations.

At first sight comparison of Eqs. (19.19) and (19.52) suggests a factor of two for the ratio of melting point and adiabatic gradients, but it is somewhat less than this because the variations of density with pressure for the two equations are represented by different moduli. This point is elaborated in the following section.

Equation (19.50) or (19.52) can be applied to iron under core conditions but are not satisfactory for extrapolation from zero pressure. The assumptions in the derivation break down in the vicinity of a triple point and there are two (at least) on the iron melting curve. The low pressure δ (body centred cubic) structure converts first to γ (face centred cubic) and then to ε (hexagonal close packed) or to something very similar. As understood in the dislocation theory of melting, the structure of a melt in equilibrium with solid crystals is a dislocated version of the crystal structure. Under conditions remote from triple points this is unambiguous, but in the vicinity of a triple point there is a progressive transition in liquid structure from a dislocated low pressure form to a dislocated high pressure form. It is not a sharp transition, as in the corresponding solid–solid phase transition, but smeared over a range in pressure and varying also with temperature, as demonstrated by Sanloup *et al.* (2000) for liquid iron in the vicinity of its δ–γ–liquid triple point.

The melting point of iron at inner core boundary (ICB) pressure is experimentally inaccessible. It is beyond the range of diamond anvils and shock waves cause melting before that pressure is reached. However, observations of the melting point of iron in its ε phase can be extrapolated to ICB pressure because no further phase transitions are involved. There is still a problem that diamond anvil measurements (e.g. Boehler, 1993) have given consistently lower estimates of T_{M} than shock waves, but in view of difficulties with both temperature measurement and pressure calibration in shock wave experiments, we extrapolate the diamond anvil results to ICB pressure by Eq. (19.52), obtaining $T_{\mathrm{M}} = 5750\,\mathrm{K}$ for iron. Depression of the melting point by solutes is also uncertain, but we allow 750 K, giving 5000 K as the ICB temperature. This is a conveniently rounded number that acknowledges the uncertainty, which is about 500 K.

The density increment by freezing is also of geophysical interest. The conventional approach to 'simple' melting, as discussed by Poirier (2000), gives an entropy increment for n moles:

$$\Delta S = nR\ln 2 + \alpha K_T \Delta V, \tag{19.53}$$

which can be used with Eqs. (19.40) and (19.50) to give

$$\Delta V = (2\gamma T_{\mathrm{M}}/K_T)nR\ln 2. \tag{19.54}$$

Applied to the core at ICB conditions this gives a density increment of $200\,\mathrm{kg\,m^{-3}}$ (1.6%). The difference between this and the observed $820\,\mathrm{kg\,m^{-3}}$ density contrast (Masters and Gubbins, 2003) is a measure of the compositional difference and the consequent gravitational energy release by progressive freezing of the inner core (Section 22.6).

19.5 Adiabatic and melting point gradients

Global geological processes at all levels are controlled or caused by convection, either thermal convection, as in the mantle, or with a strong compositional effect, as in the core. In the case of thermal convection, heat sources must maintain a temperature gradient steeper than the adiabat, even if only slightly so, and mechanical stirring, as by compositional convection, also establishes or maintains an adiabatic gradient. The existence of adiabatic gradients is an essential feature of an active planet.

To see why thermal convection can occur only in a medium with a temperature gradient exceeding the adiabatic value, consider a homogeneous medium with an arbitrary temperature gradient in which a small volume of material is displaced vertically upwards without allowing any heat transfer to or from it. Its temperature falls by adiabatic decompression as it rises. If it is then at the same temperature as its new surroundings the medium has an adiabatic gradient. If the gradient in the medium is less than this then the element is cooler and therefore denser than its new surroundings and tends to sink back to its original level. The medium is stable, with no tendency to convect. On the other hand, if the gradient in the medium is steeper than the adiabat, then the displaced volume element is hotter and less dense than its new surroundings and so tends to rise further. The medium is then unstable and may convect spontaneously. The steeper is the temperature gradient, the stronger is the convection. Since hot, rising material is displaced by cooler falling material there is a net upward transfer of heat, and if the heat sources are not maintained the temperature gradient falls and convection ceases when it reaches the adiabatic value. In the core the temperature gradient is believed to be very close indeed to adiabatic and the PREM model of the lower mantle indicates that it is only slightly steeper there (Section 17.5).

The adiabatic gradient can be written in terms of Eq. (19.19),

$$(\partial T\partial z)_{\mathrm{Adiabatic}} = (\partial T\partial P)_S\,\mathrm{d}P/\mathrm{d}z = (\gamma T/K_S)\rho g, \tag{19.55}$$

where g is gravity. With substitution for γ by Eq. (19.1) we have an alternative form in terms of α,

$$(\partial T\partial z)_{\mathrm{Adiabatic}} = \alpha Tg/C_P, \tag{19.56}$$

but, for the purpose of obtaining temperature differences, integration by Eq. (19.55) or (19.19) is more convenient. This is particularly obvious if constant γ is a sufficient approximation, so that Eq. (19.22) can be used. As pointed out in Chapter 22, it is the adiabatic temperature ratios between heat sources and sinks in a convecting medium, calculated by Eq. (19.55) or (19.56), that determine the thermodynamic efficiency of convection and hence the mechanical power that it generates. Absolute temperatures or temperature differences are not important. However, these equations are valid only for a homogeneous region and integration through phase transitions requires additional information (Section 22.3).

One reason for presenting the summary of thermodynamic relationships in Appendix E is that adiabatic gradients and adiabatic variations of physical properties are central to deep Earth physics, but most conventional treatments of thermodynamics emphasize isothermal variations. For some properties adiabatic and isothermal derivatives are very different and this becomes increasingly true as one takes progressively higher derivatives. An important example is the difference between the temperature variations of the adiabatic and isothermal bulk moduli, K_S and K_T, as represented by $\delta_S = (1/\alpha)(\partial \ln K_S/\partial T)_P$ and its more commonly quoted isothermal analogue, $\delta_T = (1/\alpha)(\partial \ln K_T/\partial T)_P$. δ_T exceeds δ_S by a factor that varies from 1.6 to 2.0 in the lower mantle. In this case it is crucial to assessment of the effect of temperature on seismic velocities, as well as to the calculation of γ by Eq. (19.33), that the correct derivative be used.

Another reason for interest in the adiabatic temperature gradient is its relationship to the melting point gradient. As appears to have been recognized first by King (1893), melting points normally increase more rapidly with pressure that the adiabatic temperature rise. King was endeavouring to rationalize the observation that the tidal rigidity of the Earth demanded solidity to great depth whereas Kelvin's pre-radioactivity calculation of the cooling of the Earth (Section 4.2) suggested that the melting point would be reached at a depth no greater than a few tens of kilometres. King concluded that the Earth would have solidified from the inside outwards, with the latent heat carried upwards convectively by a fluid layer, a process that we now understand to be occurring in the core (Sections 22.5 and 23.6). The ratio of adiabatic and melting point gradients is made quantitative by comparing Eqs. (19.19) and (19.52), using corresponding bulk moduli, given by Eqs. (19.51) and (E.1) of Appendix E, so that at the melting point

$$(\mathrm{d}T_\mathrm{M}/\mathrm{d}z)/(\partial T/\partial z)_S = 2(1+\gamma\alpha T_\mathrm{M})/(1+2\,\gamma\alpha T_\mathrm{M}), \tag{19.57}$$

which can never be less than unity. King's (1893) conclusion is a general one and all compositionally homogeneous layers in planets and satellites must solidify from the inside outwards.

In Section 10.6 we note the role of T/T_M, termed homologous temperature, in controlling the rheological properties of the mantle. It appears that sudden failure in earthquakes is restricted to regions where this ratio is less than 0.5 to 0.6. The inference is that, apart from a shallow surface layer, the only parts of the mantle where T/T_M is below this limit are the subducting slabs, and then only down to about 700 km. If this is the reason for the complete absence of earthquakes in the lower mantle, then we can use the temperature profile in Table G.2 (Appendix G) to impose limits on the solidus temperature, T_M. Adiabatic extrapolation from the 660 km phase transition gives a temperature near to the base of the mantle of about 2750 K, ignoring the steep temperature gradient right at the bottom. On this basis, T_M near to the bottom of the mantle is less than about 4580 K. Our estimate of the core–mantle boundary temperature is 3740 K, and in Section 23.5 we suggest that it was only about 200 K hotter, that is 3940 K, when the mantle had fully solidified. These numbers are compatible if we assume that, like the core, the mantle solidified from the inside outwards, with a melting point gradient steeper than the adiabat. But we note the evidence for partial melt in ultra-low velocity zones (ULVZs) at the base of the mantle (Section 17.6). The surface layer of the core is much too nearly isothermal to admit hot patches, so the ULVZs imply compositional heterogeneity. This could mean either patches of some low melting point material at the base of the mantle or pockets of the post-perovskite phase that has absorbed iron from the core (Mao *et al.*, 2006).

19.6 Thermal conduction

Through most of the Earth temperature gradients are close to adiabatic (Eqs. (19.55), (19.56)). There is, therefore, a steady flux of conducted heat at all levels. In the mantle this is only about 3% of the total heat flux, most of which is convective, except in the thermal boundary layers at the bottom and top. Conduction is important at the base of the mantle, in layer D″, where it transfers core heat into the mantle, and at the surface, where heat diffuses from the lithosphere into the atmosphere and oceans. It needs to be considered also in the transition zone, where there are temperature changes in the material convectively transported through mantle phase transitions (Sections 22.3 and 22.5) and in the diffusion of heat into cool, subducting lithospheric slabs. Thermal conductivity in the core is much higher and conducted heat is an important component of the core energy budget (Sections 21.4, 22.7 and 23.5).

In the crust and uppermost mantle the effect of pressure on conductivity is slight enough to neglect, so that measured conductivities of familiar rocks and minerals provide a reasonable estimate of the conductivity of the lithosphere. Laboratory measurements give a range of values for different rocks and minerals. We take $\kappa = 4.0\,\mathrm{W\,m^{-1}K^{-1}}$ from measurements on minerals in ultramafic rocks to be representative of the uppermost mantle (Clauser and Huenges, 1995), but only $2.5\,\mathrm{Wm^{-1}K^{-1}}$ for basaltic oceanic crust. For the deep mantle, theory takes over and it is very insecure. Lattice conduction, that is heat transport by phonons or quantized lattice vibrations, is controlled by several phonon scattering mechanisms which depend on temperature and pressure in different ways. Greatest attention in the literature has been given to phonon–phonon scattering, because this is amenable to theoretical analysis. However, while it may be the process controlling conduction in large, perfect crystals, it gives estimates of conductivity much higher than observed in ordinary materials. Phonons are scattered also by crystal imperfections of all kinds and this is the dominant process controlling conduction in minerals. Glasses are an extreme case of imperfect crystals, having dislocated liquid-type structures. As Kieffer *et al.* (1976) demonstrated in the case of fused quartz, conductivity is very much lower than for single quartz crystals and both the temperature and pressure effects are reversed. Minerals are rarely simple, but are non-stoichimetric solid solutions with a variety of atoms. In such materials phonon scattering is not observed to depend in any regular way on temperature or pressure. In the absence of contrary information we assume that lattice conductivity has no strong variation with depth in the mantle.

We have a rough check on this assumption from the thickness of the D″ layer. This layer is hottest and therefore least viscous at the bottom, so that the softened material is skimmed off into buoyant convective plumes, with the bulk of the mantle gradually collapsing on to the core to replace it. If the temperature of the plume material when it first starts rising is $\Delta T = 1000\,\mathrm{K}$ higher than the temperature of the surrounding mantle and the heat flux from the core is $\mathrm{d}Q/\mathrm{d}T = 3.5 \times 10^{12}\,\mathrm{W}$, as estimated in Chapter 21, then, with heat capacity per unit volume $(\rho C_P) = 6.6 \times 10^6\,\mathrm{J\,K^{-1}\,m^{-3}}$, the rate of removal of material is

$$\mathrm{d}V/\mathrm{d}T = (\mathrm{d}Q/\mathrm{d}T)(\rho C_P \Delta T) = 530\,\mathrm{m^3\,s^{-1}}. \qquad (19.58)$$

This material is removed from the crypto-oceanic areas of D″ (see Fig. 12.3). We have no precise estimate of the fraction of the core–mantle boundary that this represents, but the average rate of collapse of the mantle on to the core is $v = 3.5 \times 10^{-12}\,\mathrm{m\,s^{-1}}$ (0.11 mm/year). This is analogous to the ablation of meteorites and spacecraft entering the atmosphere, with heat diffusing inwards but the surface temperature maintained as material is removed. It gives an exponential temperature profile with height h above the boundary (Stacey and Loper, 1983), varying as $\mathrm{e}^{-h/H}$, where the scale height is

$$H = \eta/v \qquad (19.59)$$

and $\eta = \kappa/\rho C_P$ is the thermal diffusivity. There is a trade-off between conductivity and assumed boundary layer thickness: $\kappa/H = 2.3 \times 10^{-5}\,\mathrm{Wm^{-2}K^{-1}}$, so that, if $H = 200\,\mathrm{km}$, then $\kappa = 4.6\,\mathrm{Wm^{-1}\,K^{-1}}$. This is marginally below the lower mantle range, 5 to $12\,\mathrm{Wm^{-1}\,K^{-1}}$, inferred by Manga and Jeanloz (1997) from high pressure measurements on MgO and Al_2O_3. We can note that, with this conductivity and the boundary temperature gradient of 5 K/km (1000 K over a profile with a scale height of 200 km), the heat flow into the mantle is 3.5×10^{12} W, as adopted above. Although the numbers applied in this calculation are uncertain, they are mutually consistent, so, on this basis, we have no reason for supposing that the deep mantle conductivity is significantly different from the lithospheric value.

Another phenomenon that introduces a measure of doubt about deep mantle conduction is the radiative transfer of heat. This process occurs in stellar interiors and was introduced to geophysics by Clark (1957). If mantle minerals are moderately transparent in the near infra-red then radiation gives a contribution, κ_R, to conductivity which, in the simple case of a 'grey body', that is, one with opacity, ε, independent of wavelength, is (as Clark showed)

$$\mathrm{K_R} = (16/3)n^2\sigma T^3/\varepsilon, \tag{19.60}$$

where n is refractive index and $\sigma = 5.67 \times 10^{-8}\,\mathrm{Wm^{-2}K^{-4}}$ is the Stefan–Boltzmann constant. ε is defined as causing radiation intensity to decrease with distance x from a source as $e^{-\varepsilon x}$, so that the unit of ε is $\mathrm{m^{-1}}$. It is important in the lower mantle if $\varepsilon < 10^4\,\mathrm{m^{-1}}$, which represents a moderate transparency. However, iron-bearing minerals, typical of the lower mantle, are opaque because iron ions give strong absorption bands in the infra-red as well as optical ranges. There have been conflicting opinions about whether these absorption bands are shifted sufficiently by pressure to open a window of transparency in the lower mantle and cause significant radiative heat transfer, but Goncharov *et al.* (2006) reported that conversion of the iron ions in (Mg,Fe)O to a 'low spin' state (with electron spin moments paired) increased the mineral opacity at lower mantle pressures.

Core conductivity is dominated by electrons, not phonons. For a liquid, such as the outer core, the structural disorder gives very strong phonon scattering, as noted above, and lattice conduction is not important. Stacey and Anderson (2001) suggested a value of $3.1\,\mathrm{W\,m^{-1}K^{-1}}$, about 10% of total conductivity. Calculation of the much stronger electron contribution, κ_e, relies on the Wiedemann–Franz relationship between κ_e and the electrical conductivity, σ_e,

$$\kappa_e = L\sigma_e T, \tag{19.61}$$

where $L = (\pi k/e)^2/3 = 2.443 \times 10^{-8}\,\mathrm{W\,\Omega\,K^{-2}}$ is known as the Lorenz number, k is Boltzmann's constant and e is the electron charge. Kittel (1971) discussed the reason for the constant value of L, the same for all metals, which is that, except at very low temperatures, the scattering of electrons by phonons or by lattice defects does not depend on whether they are accelerated by an electric field or are carrying heat down a temperature gradient.

The electrical conductivity of the core is discussed in Section 24.4 with the conclusion that it varies from $2.76 \times 10^5\,\Omega^{-1}\mathrm{m^{-1}}$ at the core–mantle boundary to $2.15 \times 10^5\,\Omega^{-1}\mathrm{m^{-1}}$ just above the inner core and $2.42 \times 10^5\,\Omega^{-1}\mathrm{m^{-1}}$ in the inner core. Applying Eq. (19.61) to these values and adding $3.1\,\mathrm{W\,m^{-1}K^{-1}}$ of lattice conductivity, we have $\kappa = 28.3\,\mathrm{W\,m^{-1}K^{-1}}$ at the top of the outer core and $29.3\,\mathrm{W\,m^{-1}K^{-1}}$ at the bottom of the outer core. These numbers are very uncertain but for the purpose of our thermal calculations we assume them to apply, with a depth variation as listed in Table G.1 (Appendix G). In the inner core we take $\kappa = 36\,\mathrm{Wm^{-1}K^{-1}}$.

19.7 Temperature dependences of elastic moduli: thermal interpretation of tomography

Equations (19.34) and (19.35) define the dimensionless parameters δ_S and ε, which are used to represent the temperature dependences of elasticities. δ_S is known as the adiabatic Anderson–Grüneisen parameter, to distinguish it from its more familiar but geophysically less useful isothermal analogue, $\delta_T = (1/\alpha K_T)(\partial \ln K_T/\partial \ln T)$. It is thermodynamically related to other familiar quantities by identities in Appendix E and is calculated from Eq. (E.4), in which it may be convenient to substitute Eq. (E.8),

$$\delta_S = K'_S - 1 + q - \gamma - C'_S = K'_S - 1 + q_S - \gamma, \tag{19.62}$$

with parameters defined in Table E.1. All of them can be derived from an equation of state, with

Eq. (19.33). The estimation of γ by this equation requires δ_S as well as K'_S so the calculation of δ_S by Eq. (19.62) is iterative, but this is not a problem because the δ_S term in Eq. (19.37) is much smaller than K'_S. Calculation of q requires differentiation of Eq. (19.33) and so assumes knowledge of K'', that is the next derivative of the equation of state. This emphasizes the point that for a useful estimate of δ_S it is essential to use an equation of state for which the derivatives satisfy the relevant thermodynamic constraints. This restricts the choice to Eqs. (18.28), (18.39) or (18.45), with their integral and derivative forms.

There is no thermodynamic relationship for ε (Eq. (19.35)), corresponding to the identity forδ_S (Eq. (19.62)). Without this constraint ε cannot be determined with the same reliability as δ_S. We appeal instead to Eq. (18.67). The numerical values of the coefficients in Eq. (18.68) are specific to the lower mantle adiabat, but Eq. (18.67) must be equally valid for any adiabat, with an adjustment to $(\mu/K)_0$ but the same values of $(\mu/K)_\infty$ and K'_∞, as explained by Stacey and Davis (2004, Section 12). A simple graph (Fig. 19.3) shows how the difference in the ratio (μ/K) between two adiabats, A–C and B–C, varies with normalized pressure, P/K. From the similar triangles, ABC and DEC, we have, with Eq. (18.44),

$$\begin{aligned}\Delta(\mu/K) &= \Delta(\mu/K)_0[(P/K)_\infty - (P/K)]/(P/K)_\infty \\ &= \Delta(\mu/K)_0(1 - K'_\infty P/K). \qquad (19.63)\end{aligned}$$

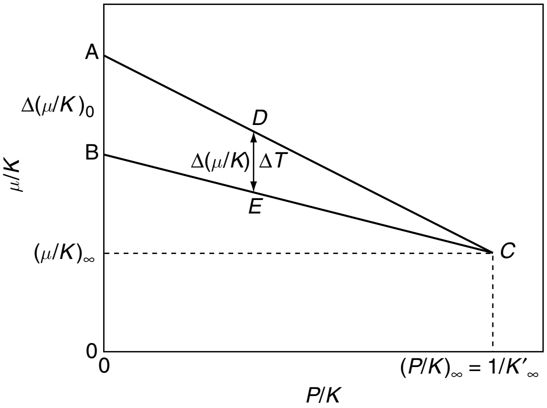


FIGURE 19.3 The ratio of elastic moduli, μ/K, plotted as a function of P/K for two close adiabats. The difference, $\Delta(\mu/K)$, decreases linearly with P/K, allowing its temperature dependence to be calculated from the zero pressure variation, $d(\mu/K)_0/dT_0$.

Since AC and BC are adiabats, the entropy difference between any pair of points, such as D–E, is the same as at A–B and can be related to the temperature difference between the adiabats at any value of P/K, allowing the variation of (μ/K) with temperature to be calculated from its variation at $P = 0$. This means that a laboratory observation of the variation of $(\mu/K)_0$ with T_0 can be combined with a calculation of the temperature dependence of K at arbitrary pressure, using δ_S (Eq. (19.62)), to give the temperature dependence of μ. The resulting equation, derived by Stacey and Davis (2004), is

$$\left(\frac{\partial \ln \mu}{\partial T}\right)_P = \frac{(\mu/K)_0}{(\mu/K)}\left[\left(\frac{\partial \ln K}{\partial T}\right)_P + \frac{d\ln(\mu/K)_0}{dT_0}\frac{T_0}{T}\frac{C_P}{C_{P_0}}\left(1 - K'_\infty \frac{P}{K}\right)\right]. \qquad (19.64)$$

A cautionary note must be applied to the use of Eq. (19.64). There are two distinct mechanisms that contribute to the temperature dependence of μ. Equation (18.67) is derived from an argument about bond forces and it leads directly to Eq. (19.64), but takes no account of an anelastic effect that also contributes. The appeal to a laboratory value of $d(\mu/K)_0/dT_0$ calibrates Eq. (19.64) by the total effect, which includes a contribution by anelasticity, unrelated to the derivation of the equation. The problem was first drawn to attention by Karato (1993), who pointed out that anelastic relaxation mechanisms that are thermally activated cause a decrease in shear modulus with temperature. Attenuation of seismic waves is caused by a time or phase lag between stress and strain (Section 10.5), which can have several causes. Some of them are thermally activated phenomena in which atomic displacements must overcome potential barriers, for which they must wait for suitable thermal impulses (phonons). The probability of a big enough impulse arriving in any time interval is represented by a frequency of occurrence

$$\nu = \nu_0 \exp(-E/kT), \qquad (19.65)$$

where E is the barrier energy, k is Boltzmann's constant and ν_0 is the 'attempt frequency' of the process, typically 10^8 to 10^9 Hz. Since this frequency (or probability) increases with temperature, the modulus decreases by an amount that can be expressed in terms of the corresponding damping of elastic waves. The effect is analysed in terms of the effect on shear wave speed (details of the algebra are given by Stacey and Davis, 2004, Appendix B), but can be expressed in terms of the effect on μ. For a seismic shear wave of frequency f,

$$(\partial \ln \mu / \partial \ln T)_{\text{Anelastic}} = -(2\pi Q_S)\ln(\nu_0/f), \qquad (19.66)$$

where Q is defined by Eq. (10.14) or (10.18) and subscript S refers to shear waves. There is no appreciable anelastic contribution to the temperature dependence of K.

Equation (19.66) gives the maximum possible effect because it assumes that all of the mechanisms that contribute to damping are thermally activated and also that Q is independent of wave frequency. An increase in Q with frequency gives a smaller effect. The maximum effect, given by Eq. (19.66), is 20% of the total $(\partial\mu/\partial T)_P$ and probably 13% would be close to the real contribution, but this is significant. In allowing for it we note that there is a general increase in Q_S as well as T with depth in the mantle, so that by Eq. (19.66) $(\partial\ln\mu/\partial T)_{\text{Anelastic}}$ decreases with depth as $1/Q_S T$. This is broadly similar to the decrease with depth of the total $(\partial \ln \mu/\partial T)_P$ by Eq. (19.64). Thus the calibration of this equation by laboratory observations of $(\mathrm{d}(\mu/K)/\mathrm{d}T)_0$ introduces no serious doubt about the calculation of ε.

Values of δ_S and ε calculated by Eqs. (19.62) and (19.64), with the definition $\varepsilon = (-1/\alpha)(\partial \ln \mu/\partial T)_P$, are listed in Appendix G for the lower mantle. They are used to calculate the thermal contribution to seismic velocity variations, for comparison with tomographic observations. Taking the equations for P- and S-wave speeds (Chapter 16),

$$V_S = (\mu/\rho)^{1/2};\ V_P = \{[K + (4/3)\mu]/\rho\}^{1/2}, \qquad (19.67)$$

and differentiating with respect to T,

$$(\partial \ln V_S/\partial T)_P = -(\alpha/2)(\varepsilon - 1), \qquad (19.68)$$

$$(\partial \ln V_P/\partial T)_P = -(\alpha/2)[K_S(\delta_S - 1) + (4/3)\mu(\varepsilon - 1)]/[K_S + (4/3)\mu]. \qquad (19.69)$$

Table 19.1 Thermal variations of seismic velocities in the lower mantle, with an extrapolation to zero pressure

R (km)	$(\partial\ln V_S/\partial T)_P$ ($10^{-5}\mathrm{K}^{-1}$)	$(\partial\ln V_P/\partial T)_P$ ($10^{-5}\mathrm{K}^{-1}$)	$(\partial\ln V_S/\partial\ln V_P)_P$	$(\partial\ln V_\phi/\partial\ln V_S)_P$
3480	− 2.473	− 1.001	2.470	0.0492
3600	− 2.561	− 1.050	2.440	0.0544
3630	− 2.584	− 1.062	2.432	0.0574
3800	− 2.714	− 1.136	2.388	0.0636
4000	− 2.870	− 1.229	2.335	0.0739
4200	− 3.042	− 1.334	2.280	0.0852
4400	− 3.228	− 1.449	2.228	0.0967
4600	− 3.451	− 1.595	2.164	0.1117
4800	− 3.703	− 1.762	2.102	0.1276
5000	− 3.990	− 1.960	2.036	0.1460
5200	− 4.341	− 2.209	1.965	0.1677
5400	− 4.760	− 2.518	1.890	0.1934
5600	− 5.283	− 2.922	1.808	0.2254
5701	− 5.543	− 3.154	1.758	0.2474
$P = 0$	− 8.796	− 6.080	1.447	0.4315

We are interested also in the hydrodynamic or 'bulk sound' velocity

$$V_\phi = (V_P^2 - (4/3)V_S^2)^{1/2} = (K_S/\rho)^{1/2}, \qquad (19.70)$$

for which

$$(\partial \ln V_\phi/\partial T)_P = -(\alpha/2)(\delta_S - 1). \qquad (19.71)$$

This is not directly observable but is calculated from V_P and V_S. The significance is that its temperature variation does not depend on ε but only on the better determined δ_S. Temperature variations of the lower mantle velocities are listed in Table 19.1.

Velocity heterogeneities are observed at all depths in the mantle by seismic tomography (Section 17.7). They are generally stronger in the upper mantle, where subducting lithospheric slabs present strong contrasts in temperature-dependent properties. The inference is that in the lower mantle also the heterogeneities reflect tectonics, present and past, with high velocities marking cooler than average material. However, we can use the calculations of δ_S and ε to show that temperature cannot be the only cause and that compositional variations must be invoked. Particular attention is given to ratios of the velocity variations in Eqs. (19.68), (19.69) and (19.71), for which knowledge of α is not required,

$$(\partial \ln V_S/\partial \ln V_P)_P = [1 + (4/3)(\mu/K_S)]/[(\delta_S - 1)/(\varepsilon - 1) + (4/3)(\mu/K_S], \qquad (19.72)$$

$$(\partial \ln V_\phi/\partial \ln V_S)_P = (\delta_S - 1)/(\varepsilon - 1). \qquad (19.73)$$

In these equations we see two consequences of the fact that δ_S decreases with depth, approaching unity deep in the lower mantle, as in the listing in Appendix G. By Eq. (19.73), $(\partial \ln V_\varphi/\partial \ln V_S)$ becomes very small, but it remains positive because, with these calculations, δ_S does not fall below 1. Two groups reported negative correlations between V_φ and V_S (Robertson and Woodhouse, 1996a,b; Su and Dziewonski, 1997). A thermal explanation would require $\delta_S < 1$, contrary to calculations using Eq. (19.62) (because $\varepsilon > 1$ is not in doubt). This is evidence of compositional heterogeneity (see Section 17.8), a conclusion that does not depend on knowledge of ε. The very small value of $(\delta_S - 1)/(\varepsilon - 1)$ deep in the lower mantle means also that, by Eq. (19.72), $(\partial \ln V_S/\partial \ln V_P)_P$ is insensitive to the value of ε. The observed variation in this ratio with depth is broadly similar to the calculated variation (Fig. 19.4), indicating that temperature variations account for a major part of it, although not all. Unfortunately, insensitivity of critical observations to the value of ε prevents an observational check on the magnitude of the anelastic effect.

19.8 Anharmonicity

The theory of specific heat, based on the principle that lattice modes are harmonic oscillators (Section 19.2), and the interpretation of thermal expansion by Grüneisen's theory (Section 19.3) assume that if a material is held at constant volume then the frequencies of its lattice modes are independent of temperature or vibration amplitude. This requires the oscillations to be sinusoidal (harmonic) and that the atoms move under the influence of interatomic forces proportional to their displacements from equilibrium, with potential energies proportional to squares of the displacements. Although this is a useful first approximation, as Fig. 18.4 illustrates, it misses some of the important physics and we need to consider the consequences of departures from the harmonic assumption. Thermal expansion and the strong pressure dependence of bulk modulus are consequences of the asymmetry of atomic potentials, as in Fig. 18.4, and are therefore anharmonic effects. However, solid state and mineral physicists have generally made a distinction between these two effects and other consequences of anharmonicity, such as departures from the Dulong–Petit theory of specific heat (Section 19.2). Acceptance of thermal expansion and pressure-dependent elasticity, with neglect of temperature variations of other thermal properties, such as specific heat, is referred to as the quasi-harmonic approximation (QHA), for reasons that are discussed below. The adequacy of QHA has been vigorously debated by the mineral physics community. This section is concerned with the physical principles, following the argument of Stacey and Isaak (2003), who interpreted the distinction

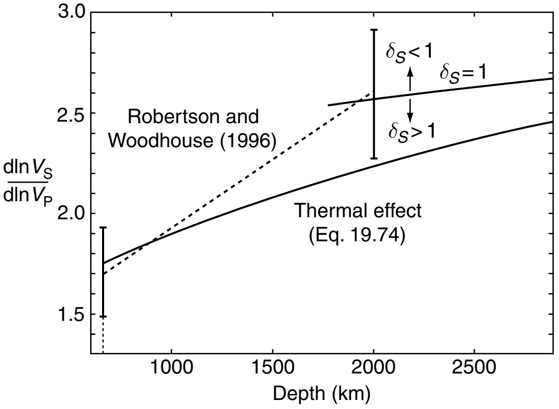


FIGURE 19.4 A comparison of the ratio of P-wave and S-wave velocity variations in the lower mantle with Eq. (19.72), using values of δ_S and ε tabulated in Appendix G.

between QHA and full anharmonicity in terms of derivatives of the atomic potential function (ϕ in Fig. 18.4). QHA can be explained by non-zero $d^3\phi/dr^3$ with neglect of the higher derivatives that are required for temperature-dependent specific heat, etc. Stacey and Isaak termed this first-order anharmonicity. The neglected higher-order anharmonic effects are conventionally referred to simply as anharmonicity.

The quasi-harmonic/anharmonic distinction is blurred by the fact that there are two different mechanisms involved. The asymmetry of atomic potential functions referred to above causes what we call type 1 or bond anharmonicity. This occurs in all situations and is the only type of anharmonicity in solids in which all atoms have neighbours in opposite pairs. Asymmetry of crystal structures, with bonds that do not occur in opposite pairs, causes type 2 or structural anharmonicity. In this case atomic oscillations would be anharmonic even if individual atomic bonds were perfectly harmonic. Then the two types are superimposed and they have effects that are generally opposite in sign.

Consider first type 1 anharmonicity arising from a linear oscillation of an atom between two neighbours, as in Fig. 19.5(a). This is the situation observed in minerals such as MgO, which has cubic symmetry. It is a simplification that ignores motion in perpendicular directions and correlated motion of neighbours but it suffices for a consideration of the effect of bond asymmetry on atomic motion. For small displacements, x, of atom B from its equilibrium position between the fixed atoms A and C, we can write the potential energies of the two bonds as Taylor expansions about the equilibrium atomic spacing, a,

$$\phi(a \pm x) = \phi(\mathrm{a}) \pm \phi'(a)x + (1/2!)\phi''(a)x^2 \\ \pm (1/3!)\phi'''(a)x^3 + (1/4!)\phi^{\mathrm{iv}}(a)x^4 + \cdots. \tag{19.74}$$

The total energy is the sum of the two potentials, with subtraction of the equilibrium energy at $x = 0$,

$$E = \phi(a + x) + \phi(a - x) - 2\phi(a) \\ = \phi''(a)x^2 + (1/12)\phi^{\mathrm{iv}}(a)x^4 + \cdots. \tag{19.75}$$

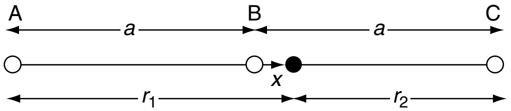


FIGURE 19.5(a) Linear oscillation of an atom, B, between neighbours A and C that are assumed fixed.

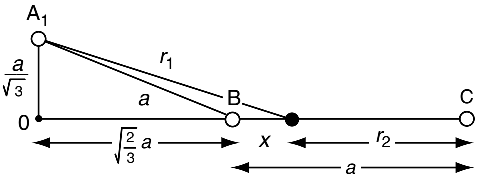


FIGURE 19.5(b) Oscillation of an atom, B, in a crystal with the diamond structure, having the four bonds to each atom, directed to the corners of a tetrahedron. B is displaced from equilibrium by a small distance, x, on the line OBC, where O is the mid-point of an equilateral triangle formed by neighbours A_1, A_2, A_3, all equally spaced from one another and from B.

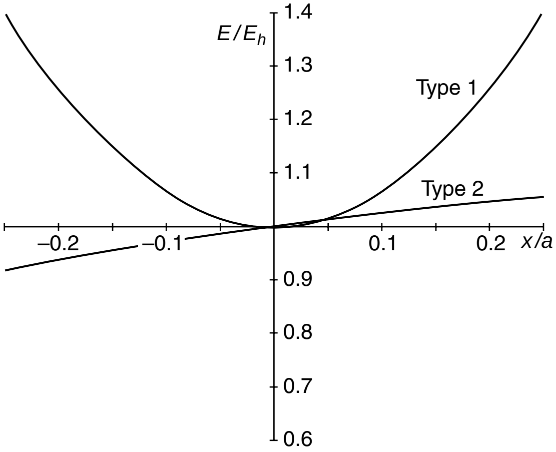


FIGURE 19.6 Energies of atomic displacement from equilibrium for geometries in Fig. 19.5(a), giving type 1 (bond) anharmonicity (Eq. (19.76)), and Fig. 19.5(b), giving type 2 (structural) anharmonicity (Eq. (19.82)).

Odd powers of x and odd derivatives of ϕ are eliminated by the symmetry. For very small x this reduces to the energy for harmonic bonding, $E_h = \phi''(a)x^2$, so that the departure from this situation is represented by

$$E/E_h = 1 + (\phi^{\mathrm{iv}}(a)/\phi''(a))x^2/12. \tag{19.76}$$

In Fig. 19.6 this is plotted for the Born–Mie potential (Eq. (18.2) with just two terms and $m = 1$, $n = 5$, to give $K_0' = 4$). For all plausible potential functions ϕ^{iv} is positive, giving an excess energy relative to harmonic bonds. It has the effect of driving atom B away from its extreme excursions, reducing the time spent at large values of x and therefore the mean potential energy of thermal vibration. A result is a slight reduction in specific heat, relative to the Dulong–Petit limit (Eq. (19.3)). Anderson and Zou (1990) observed this effect in MgO.

The geometry for calculation of type 2 anharmonicity is shown in Fig. 19.5(b) for the case of a diamond structure. To isolate the type 2 effect from a superimposed type 1 effect we assume that all four of the bonds to the oscillating atom, B, are perfectly harmonic, that is

$$\phi = \phi(a) + A(r - a)^2, \tag{19.77}$$

where A is a constant. Subtracting the energy at equilibrium ($x = 0$), the total potential energy at displacement x is

$$E = A(r_2 - a)^2 + 3A(r_1 - a)^2, \tag{19.78}$$

where r_2 is simply $(a - x)$, but

$$r_1 = [a^2 + x^2 + 2\sqrt{(2/3)}ax]^{1/2}. \tag{19.79}$$

Binomial expansion for small x gives

$$\begin{aligned}(r_1 - a)^2 = a^2[(2/3)(x/a)^2 &+ \sqrt{(2/27)}(x/a)^3 \\ &- (7/36)(x/a)^4 + \cdots].\end{aligned} \tag{19.80}$$

Substituting this in Eq. (19.78), we have

$$E/Aa^2[3(x/a)^2 + \sqrt{(2/3)}(x/a)^3 - (7/12)(x/a)^4 + \cdots] \tag{19.81}$$

and, as for the type 1 calculation, we take the ratio of total to harmonic energy,

$$E/E_h = 1 + \sqrt{(2/27)}(x/a) - (7/36)(x/a)^2. \tag{19.82}$$

The important feature of Eq. (19.82) is the linear term. The potential energy is strongly asymmetrical and so is quite different from the type 1 effect, with which it is compared in Fig. 19.6. Note that these are small departures from the parabolic (harmonic) form and we are considering $x/a \ll 1$, so that the first terms in Eqs. (19.75) and (19.81) are dominant and the anharmonicity is a small superimposed effect. For type 1 (bond) anharmonicity it depends on the precise form of the atomic potential function, but all plausible forms give a result very similar to that plotted in Fig. 19.6. As mentioned, it gives a small decrease in C_V with T, 2% to 3% for MgO at 2000 K. When type 2 (structural) anharmonicity occurs, type 1 is superimposed but is often masked by the type 2 effect. This depends on crystal structure and is most obvious for the tetrahedral bonding represented in Fig. 19.5(b) and used to derive Eq. (19.82). This is common in silicates because Si^{4+} ions favour tetrahedral bonding to surrounding O^{2-} ions. It allows atoms to vibrate towards ‘soft’ gaps in the crystal structure, with correspondingly greater amplitudes, causing C_V to increase with temperature, as observed in forsterite (Mg_2SiO_4) by Gillet *et al.* (1991). Another consequence of ‘soft’ vibrations is a lowered thermal expansion coefficient, as considered below. This is particularly obvious in crystals of Si and Ge, which have negative coefficients over the limited temperature ranges for which asymmetrically soft modes are most significant, but the effect is quite general and common (low pressure) silicates have low expansion coefficients.

Although we tend to think of the thermal expansion coefficient, α, and in particular its temperature dependence, as indicative of anharmonicity, this is not strictly correct. α is temperature dependent even with the quasi-harmonic assumption. The true indicator is the Grüneisen parameter, γ (Eq. (19.1)). For α we have a thermodynamic identity,

$$(\partial\alpha/\partial T)_V = \alpha^2[q - 1 - (\partial \ln C_V/\partial \ln V)_T(1 + 1/\gamma\alpha T)], \tag{19.83}$$

which does not vanish with the quasi-harmonic assumption (because $q = (\partial \ln \gamma/\partial \ln V)_T \neq 1$), even with the classical assumption that C_V is constant. On the other hand, as in Eq. (19.28),

$$(\partial\gamma/\partial T)_V = (1/T)(\partial \ln C_V/\partial \ln V)_S. \tag{19.84}$$

So, we examine the anharmonicity of γ resulting from the two bond geometries considered above. The Mie–Grüneisen equation (Eq. (19.21)) can be taken to define γ as the ratio of thermal pressure to thermal energy per unit volume. The effect on thermal energy of the two types of anharmonicity is represented by the energies of atomic displacement in Eqs. (19.76) and (19.82). We now seek expressions for the bond forces that cause thermal pressure.

For the type 1, symmetrical situation in Fig. 19.5(a), we need to calculate the average of the forces exerted by B on A and C. It is important to note that we cannot do this by differentiating Eq. (19.75) because that gives the net force on B, that is the difference between the two bond forces, not the sum of them. We write Taylor expansions for the individual bond forces in the same way as Eq. (19.74) expands the energies, noting that bond force is simply $-\phi'$:

$$\phi'(r) = \phi'(a) \pm \phi''(a)x + (1/2!)\phi'''(a)x^2 \pm (1/3!)\phi^{iv}(a)x^3 + (1/4!)\phi^{v}(a)x^4. \tag{19.85}$$

Subtracting the non-thermal force, $\phi'(a)$, the sum of the two forces in Eq. (19.85) is

$$F = \phi'''(a)x^2 + (1/12)\phi^{v}(a)x^4. \tag{19.86}$$

Considering x to be the instantaneous displacement in thermal vibration, we can take averages of Eqs. (19.75) and (19.86) and then the ratio of these equations is the ratio of thermal pressure to thermal energy, that is γ. We have ignored geometrical factors required for integration over all bond orientations and so write this as a proportionality:

$$\begin{aligned}\gamma \propto \frac{\langle F\rangle}{\langle E\rangle} &= \frac{\phi'''(\mathrm{a})\langle x^2\rangle + \frac{1}{12}\phi^{v}(\mathrm{a})\langle x^4\rangle}{\phi''(\mathrm{a})\langle x^2\rangle + \frac{1}{12}\phi^{iv}(\mathrm{a})\langle x^4\rangle}\\ &= \frac{\phi'''(\mathrm{a})}{\phi''(\mathrm{a})}\cdot\frac{1+[\phi^{v}(\mathrm{a})/\phi'''(\mathrm{a})]\langle x^4\rangle/12\langle x^2\rangle}{1+[\phi^{iv}(\mathrm{a})/\phi''(\mathrm{a})]\langle x^4\rangle/12\langle x^2\rangle}.\end{aligned} \tag{19.87}$$

By separating the factor ϕ'''/ϕ'' in this equation we draw attention to the distinction between the quasi-harmonic approximation (QHA) and the higher anharmonic effects. QHA accepts positive thermal expansion ($\gamma > 0$) and is represented by ϕ'''/ϕ'', but if the higher derivatives, ϕ^{iv} and ϕ^{v}, are assumed to be zero then γ is independent of vibration amplitude, or temperature. The second fraction in this equation gives the anharmonic temperature dependence. Stacey and Isaak (2003) pointed out that, for a range of commonly used potential functions and finite strain theories, the ratios ϕ^{v}/ϕ''' and ϕ^{iv}/ϕ'' are very similar. This means that the anharmonic temperature dependence of γ is weaker than that of C_V, which is represented by the denominator of Eq. (19.87). The anharmonicity of C_V is better observed and, particularly for type 1 structures, is found to be quite small. That the anharmonicity of γ is smaller still makes it insignificant for type 1 bonding, especially under pressure, which suppresses it.

In the asymmetrical, type 2 situation of Fig. 19.5(b) we sum the bond forces resolved in the direction of OBA, assuming individual bond forces to be harmonic (Eq. (19.77)), as before, so that $\phi' = 2A(r - a)$. Writing the sum of the four resolved forces and expanding binomially for small x, we have a total force

$$F = Aa[-2(x/a) - \sqrt{6}(x/a)^2 + (7/3)(x/a)^3 - (25/36)(x/a)^4 + \cdots]. \tag{19.88}$$

Now we can compare Eqs. (19.88) and (19.81) for an idea of the effect on γ of structural anharmonicity. Strictly, we need time-averaged values, which require numerical integration to handle the non-harmonic oscillation, but as an approximation we assume that positive and negative x occur equally often, so that odd powers of x drop out in the averaging process, as for Eqs. (19.75) and (19.86). Then we have

$$\gamma \approx a\frac{\langle F\rangle}{\langle E\rangle} = \frac{-\sqrt{6}\left\langle (x/a)^2\right\rangle - \frac{25}{36}\left\langle (x/a)^4\right\rangle}{3\left\langle (x/a)^2\right\rangle - \frac{7}{12}\left\langle (x/a)^4\right\rangle}. \tag{19.89}$$

With the assumed geometry γ and hence α are negative.

Equation (19.89) is not an estimate of γ for any real situation, but an indication of the fundamental reason why crystals with open structures, most particularly those with tetrahedral bonding, such as many common silicates, have low expansion coefficients, and even negative coefficients under special conditions. It is a consequence of structural or type 2 anharmonicity. A three-dimensional treatment, superposition of bond or type 1 anharmonicity and rigorous integration of atomic motion complicate the picture, but do not change it fundamentally. A conclusion from Eq. (19.89) is that structural anharmonicity allows a stronger temperature dependence of γ than does bond anharmonicity. This is seen in the ratio of the x^4 and x^2 terms. However, all anharmonic effects diminish strongly with pressure, both because vibration amplitudes become decreasing fractions of atomic spacing and because the structural effect decreases with close-packing. Although anharmonicity may appear to be a subtle effect that can be neglected for many purposes, it has a central role in phenomena such as the temperature dependences of elastic moduli.

20

The surface heat flux

20.1 Preamble

As has been known for at least 200 years, there is a steady flux of heat from the crust into the atmosphere and oceans. The crustal temperature gradient, measured in the top kilometre or so in stable continental areas, is typically 0.025 $\mathrm{K\,m^{-1}}$ (25 K/km). With a conductivity of 2.5 $\mathrm{W\,m^{-1}\,K^{-1}}$ this corresponds to a conducted heat flux of 62.5 $\mathrm{m\,W\,m^{-2}}$. The heat flux can be much higher in continental tectonic and thermal areas, but they cover a sufficiently small fraction of the total area of the Earth to have little influence on the global average. There are now more than 10 000 measurements of the heat flux from continents, with a wide range of values but sufficient data to be confident of the average, 65 $\mathrm{m\,W\,m^{-2}}$ (Pollack *et al.*, 1993).

The observation of the continental temperature gradient has had a central role in the history of geophysics. Extrapolated downwards, it suggested that rock would be at its melting point at a depth of 50 km or so. This observation prompted Kelvin's cooling Earth calculation and the age of the Earth debate in the 1800s (Section 4.2). When radioactivity was discovered and found to be concentrated in crustal rocks, especially granite, it invited the conclusion that crustal radiogenic heat explained all of the observed heat flux and that the deep Earth was thermally passive. This discovery effectively suppressed ideas about convection that had been debated for the previous 40 years but were not revived for another 60 years. It illustrates the limited perception of the geothermal flux when observations were restricted to the continents. Now we have more observations from the ocean floor, which is more youthful than the continents, and has much less radioactivity. Studies of the ocean floor, and especially its heat flux, are essential to the modern understanding of global geology, based on the concept of convection. The sea floor is the exhaust of a giant heat engine that follows familiar thermodynamic principles (Chapter 22) and provides the power for all tectonic processes.

The ocean floor is cooled by convective circulation of sea water in cracks, as well as by conduction, so that the measured temperature gradient in the sediment underestimates the heat flux. The heat loss from the ocean floor is more effectively estimated by thermal shrinkage, apparent as increasing ocean depth with age. Allowing for this, the integrated total heat flux for the Earth is 44.2×10^{12} W, an average of 87 $\mathrm{m\,W\,m^{-2}}$ (Pollack *et al.*, 1993). Relative to other surface processes this is quite a small number. Solar energy reaching the Earth is greater by a factor 2000. We can consider what it means for cooling of the Earth, for which the total heat capacity is $6.6 \times 10^{27}\,\mathrm{J\,K^{-1}}$. Ignoring heat sources, the heat flux would imply cooling at an average rate of $6.7 \times 10^{-15}\,\mathrm{K\,s^{-1}}$ (210 K/10^9 years), emphasizing that cooling of very large bodies is slow, even with convection. The real cooling rate is about one third of this as radiogenic heat provides two thirds of the heat loss.

The surface temperature of the Earth is subject to variations with a wide range of time scales. Diurnal and annual periods are obvious. They propagate downwards with exponential attenuation: for the annual cycle the scale depth is only about 3 m and the effect is of no geophysical interest. But longer term changes are seen in boreholes at depths that are used for heat flux observations, to which they add noise. But they constitute a signal that is of interest in probing climatic variations on time scales extending to thousands of years. This has become a reason for making temperature measurements in continental boreholes, semi-independently of the heat flux problem.

20.2 The ocean floor heat flux

As mentioned in Section 19.4, water has the unusual property of expanding on freezing. This is 'anticipated' by the liquid, which has a negative expansion coefficient below 4 °C, its temperature of maximum density. For sea water the maximum occurs at 2 °C. Cold polar water of maximum density sinks to the sea floor and spreads out over all the ocean basins, stabilizing the temperature there. This has made possible widespread measurements of ocean floor heat flux. Kilometre deep holes are not needed for measurements of temperature gradient; spikes of a few metres length, fitted with sensitive thermistors and dropped to the ocean floor, penetrate the soft sediment and quickly adjust to the ambient temperature gradient. With the stable temperature, a measurement of the gradient over a few metres suffices for an estimate of the heat flux. Accuracy of the estimate is limited only by the need to determine the conductivity of undisturbed sediment ($\sim 1\,\mathrm{W\,m^{-1}K^{-1}}$) and to ensure that the sediment is thick enough to smear out the effect of the rough topography of its contact with the underlying, more conductive basalt.

Ocean floor survives for no more than 200 million years and its average lifetime, before subduction, is about 90 million years (the ratio of total ocean floor area, $3.1 \times 10^8\,\mathrm{km}^2$, to its rate of production, $3.4\,\mathrm{km}^2$/year). The average age of the ocean floor is about 65 million years. It appears as fresh igneous crust at mid-ocean ridges, from which it spreads towards subduction zones, as discussed in Chapter 12. The ocean floor crust is only 7 km thick but the cooling by its exposure to the ocean water extends to a depth of 100 km or more. Formally, this defines the lithosphere, the cool upper layer of the mantle that is more rigid than the hotter rock beneath, but here we use the term more loosely to refer to the layer that eventually becomes lithosphere. The cooled layer thickens with age and shrinks thermally, causing the ocean depth to increase with lithospheric age and therefore distance from the ridge sources. Observations of decreasing heat flux and increasing depth with age rank as two of the most important foundations of the theory of global tectonics, but detailed interpretation is not entirely straightforward. Circulation of sea water through cracks, variations in thermal properties with depth and motion of the underlying asthenosphere introduce complications that still lack comprehensive explanations.

We consider first a simple model by treating the oceanic lithosphere as a diffusively cooling half space. Assuming that its physical properties and initial temperature are uniform and that it has no internal heat sources, we solve the diffusion equation in one dimension (depth z):

$$\frac{\partial T}{\partial t} = \eta \frac{\partial^2 T}{\partial z^2}. \tag{20.1}$$

This equation relates the heating or cooling of any material element to the local variation in temperature gradient. It describes the departure of the heat flow from a steady state. Here η is the thermal diffusivity, defined by

$$\eta = \kappa/\rho C_P \tag{20.2}$$

in terms of conductivity, κ, density, ρ, and specific heat, C_P. κ and η vary with temperature and pressure as well as with composition, but for the purpose of the simple half-space model we assume uniform values, corresponding either to the igneous crust ($\kappa = 2.5\,\mathrm{W\,m^{-1}K^{-1}}$, $\eta = 0.75 \times 10^{-6}\,\mathrm{m^2\,s^{-1}}$) or uppermost mantle ($\kappa = 4.0\,\mathrm{W\,m^{-1}K^{-1}}$, $\eta = 1.0 \times 10^{-6}\,\mathrm{m^2\,s^{-1}}$). Then, with the uniform initial temperature, T_0, and suddenly imposed cold surface, $T = 0$, at $z = 0$, the temperature step, $-T_0$, spreads out as it propagates downwards. At any later time, t,

$$T = T_0 \ \mathrm{erf}(z/(\eta t)^{1/2}), \ t > 0, \qquad (20.3)$$

where the error function is defined by

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-\varsigma^2) \mathrm{d}\varsigma. \qquad (20.4)$$

This function is tabulated in several handbooks, but requires care because sometimes it is erf $(x/\sqrt{2})$ that is tabulated (eg. Dwight, 1961, Table 1054). Also there are series approximations for small and large x (Dwight, 1961, Items 590–592).

The temperature gradient, obtained by differentiating Eq. (20.3), is

$$(\partial T/\partial z)_t = T_0/(\pi\eta t)^{1/2} \exp(-z^2/4\eta t). \qquad (20.5)$$

At the surface, $z = 0$,

$$(\mathrm{d}T/\mathrm{d}z)_0 = T_0/(\pi\eta t)^{1/2} \qquad (20.6)$$

and the flux of heat conducted to the sea (per unit area) is

$$\frac{\dot{Q}}{A} = \kappa \left(\frac{\mathrm{d}T}{\mathrm{d}z}\right)_0 = \frac{\kappa T_0}{\sqrt{\pi\eta t}} = T_0 \sqrt{\kappa\rho C_P/\pi} * t^{-1/2}. \qquad (20.7)$$

This is the $t^{-1/2}$ variation of heat flux with lithospheric age that is characteristic of the diffusive half space model. In Fig. 20.1 it is compared with observations, as summarized in Table 3 of Stein (1995). Each data point is an average of several hundred individual values from all oceans with an age range centred at the marked point and scatter indicated by its error bar. It is important to note that the data in Fig. 20.1 give the conducted heat flux, being the product of temperature gradient and conductivity, and that there is an additional heat transfer by the convective circulation of sea water in cracks. This sharply reduces the temperature and near surface temperature gradient, making the conducted heat much less than the diffusive half-space model suggests for young lithosphere. The almost constant conducted heat flux after about 30 million years is not as easily explained.

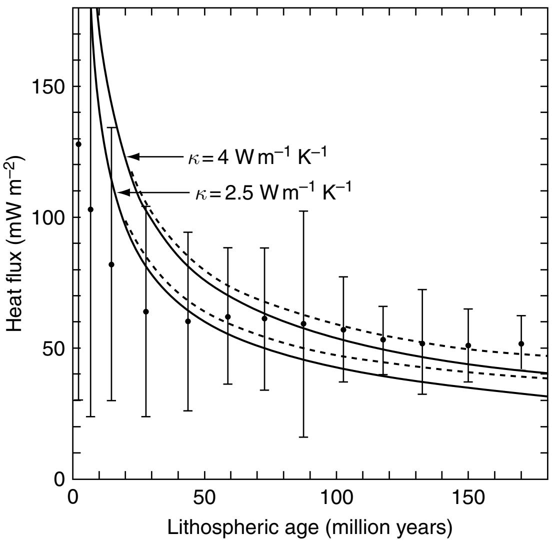


FIGURE 20.1 Variation in conducted heat flow with age of the ocean floor. Data points from Table 3 of Stein (1995) compared with the $t^{-1/2}$ variation of the cooling half space model (Eq. (20.7)) for two values of thermal conductivity. Broken lines indicate the effect of introducing radiogenic heat.

Complementary information is provided by the variation in ocean depth, assuming that it is due to thermal contraction, which is proportional to the integrated heat loss. Since this includes the hydrothermal cooling, which is otherwise difficult or impossible to estimate, it must be regarded as the definitive evidence of cooling. For a change ΔT in the temperature of any material element of mass m and volume V the ratio of thermal expansion or contraction to heat gain or loss is

$$\frac{\Delta V}{\Delta Q} = \frac{\alpha V \Delta T}{m C_P \Delta T} = \frac{\alpha}{\rho C_P} = \frac{\gamma}{K_S}, \qquad (20.8)$$

where γ is the Grüneisen parameter, defined by Eq. (19.1). This is a convenient expression to use because temperature variations in α and C_P are almost mutually cancelling and are allowed for by using γ. With the assumption of uniform properties, the total thermal contraction is proportional to the integrated heat loss from all depths. If we consider only diffusive cooling, then integration of Eq. (20.7) gives

$$\frac{\Delta Q}{A} = \frac{1}{A} \int \dot{Q} \mathrm{d}t = T_0 \sqrt{\kappa\rho C_P/\pi} \cdot 2t^{1/2}, \qquad (20.9)$$

so that, with Eq. (20.8),

$$\frac{\Delta V}{A} = (2\gamma T_0/K_S) \sqrt{\kappa\rho C_P/\pi} \cdot t^{1/2}. \qquad (20.10)$$

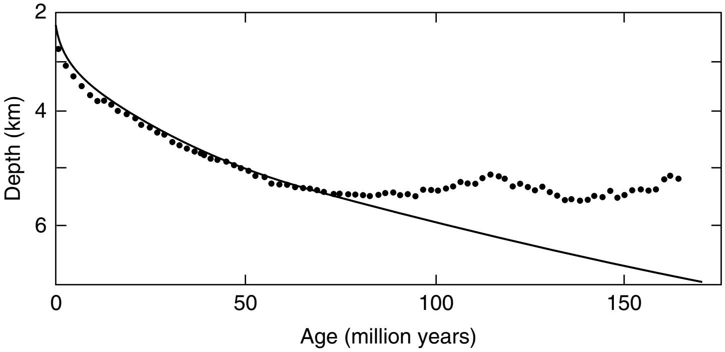


FIGURE 20.2 Ocean depth as a function of lithospheric age. Average of North Atlantic and North Pacific data as plotted by Stein and Stein (1992), compared with Eq. (20.11). A correction for radiogenic heat has been applied to the theoretical curve.

We note that, as the lithosphere shrinks, becoming denser, and the water depth increases, so an additional water load is added and this is isostatically compensated by further depression of the ocean floor. The variation in ocean depth is $(1-\rho_w/\rho_m)^{-1} = 1.437$ times the lithospheric contraction, where $\rho_w = 1025\,\mathrm{kg\,m^{-3}}$ and $\rho_m = 3370\,\mathrm{kg\,m^{-3}}$ are the water and mantle densities. Thus, by Eq. (20.10), with inclusion of the isostatic factor, we can write the increase in ocean depth with age as

$$\Delta z = (2.87\gamma T_0/K_S)(\kappa\rho C_P/\pi)^{1/2}t^{1/2}. \qquad (20.11)$$

This is plotted in Fig. 20.2 and compared with observations from the North Atlantic and North Pacific, as plotted by Stein and Stein (1992, see also Stein and Stein, 1994).

The immediate conclusion from Fig. 20.2 is that the oceans cease to become deeper after about 70 million years and this is not consistent with the cooling half-space model, even allowing for radiogenic heat in the lithosphere. Thus the observations represented by Figs. 20.1 and 20.2 both indicate deficiencies in the half-space model, prompting an alternative, referred to as the plate model. This assumes that mature lithosphere reaches a limiting thickness, after which it ceases to cool further and acts as a steady-state conducting layer, with a lower boundary maintained at a fixed temperature by contact with the asthenosphere (e.g. Stein and Stein, 1992). While this appears to explain the stabilization of both heat flux and depth, it introduces another problem. Unless it is continuously, and rapidly, replaced, the asthenosphere can provide the necessary heat only by cooling, and so adding to the lithosphere and, indeed, dispersion of Rayleigh waves with oceanic paths indicates that the lithosphere continues to thicken (Zhang and Tanimoto, 1991; Zhang and Lay, 1999; Maggi *et al.*, 2006). Thus the plate model is also unsatisfactory and we re-examine the observations for clues to effects that are not recognized in these theories.

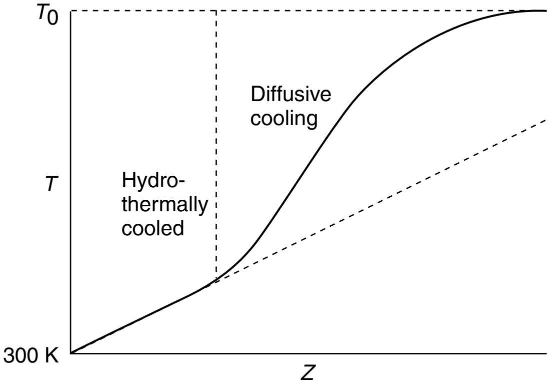


FIGURE 20.3 General form of the lithosphere temperature profile with hydrothermal cooling.

Hydrothermal cooling of young lithosphere is clearly observed. At least the upper part is cooled much more rapidly than by diffusion. Assuming that cracking is necessary for sea water circulation and that it is confined to a shallow layer, perhaps a few kilometres thick, then it establishes a temperature profile represented qualitatively by Fig. 20.3. The rapidly cooled layer would leave the deeper lithosphere still hot, with a steep temperature gradient between the layers of hydrothermal and diffusive cooling. Then, if the water is eventually cut off (perhaps by sediment choking the cracks), the temperature profile will adjust towards the diffusive one, given by Eq. (20.3). This means some reheating of the shallow lithosphere at the expense of the still hot deeper part, maintaining the temperature gradient in the shallow layer. It would also tend to offset the overall thermal contraction of the lithosphere if the thermal expansion coefficient decreases with depth. However, from an attempt to model this effect numerically we find that, although it may be a contributory cause of the discrepancies between observations and the half-space theory, it is quantitatively inadequate and further explanations are needed.

The fact that the conducted heat flux stabilizes to 50 or 60 $\mathrm{m\,W\,m^{-2}}$ after 25 or 30 million years (Fig. 20.1) means that the temperature gradient near the surface is stabilized to a constant value long before the thermal contraction, indicated by Fig. 20.2, stops, at about 70 million years. Also, the continued increase in lithospheric thickness indicated by seismology after this time disallows the supposition, central to the plate model, that a static diffusive thermal structure has been established. We appear to require not one but two further effects, because the heat flux and ocean depth are stabilized on different time scales. Possibilities are hydrothermal circulation that extends deep into the lithosphere and redistributes heat without necessarily exhausting it to the ocean, and an asthenosphere that grows thicker as the over-riding lithosphere approaches a subduction zone. We examine the arguments for each of these possibilities briefly, but a complete and satisfying explanation still eludes us.

The more or less steady conducted heat flux after about 30 million years implies a linear temperature gradient that could not be established in mature lithosphere, of order 100 km thick, by thermal diffusion with only a thin hydrothermally cooled layer. A linear temperature gradient that remains more or less constant is characteristic of fluid convection. The heat flux plotted in Fig. 20.1 is diffusive, being the product of conductivity and temperature gradient in the surface layer, but we cannot conclude that the constant temperature gradient is indicative of a diffusive layer in a steady state. The fact that the gradient reaches a constant value after 30 million years while the ocean continues to deepen for another 40 million years requires another explanation. It suggests that fluid circulation extends much deeper than usually supposed and that it lasts for the entire life of the oceanic lithosphere.

Now consider the structure of the asthenosphere underlying a moving oceanic plate. It has no sharp boundary. At the top it is welded to the lithosphere and moves with it, while the diffuse and ill-defined bottom boundary is anchored to the deeper mantle. It is a shearing layer, with roughly 50% of the material moving with the plate. Where does it go when the plate approaches a subduction zone? It is hot and buoyant and resists subduction, but it is also deformable and will tend to avoid subduction by thickening beneath the ageing lithosphere. The lithosphere itself has increasing negative buoyancy as it continues to cool and thicken, but the fraction of the underlying asthenospheric ‘toothpaste’ that avoids subduction gives increasing support, offsetting the increase in ocean depth that would otherwise be expected. In situations where the lithosphere is approaching a subduction zone that is ‘rolling back’ towards the ridge source, it would generate pressure in the lithosphere and upper mantle that would have a similar effect. Neither of these attempted explanations relates to the situation in the Atlantic Ocean, with both margins bounded by continents that are parts of the moving lithospheric plates and are receding from the central ridge. In that case we might suppose that the continents have deep roots that cause them to move less freely over a thinner or more viscous asthenosphere and are moving apart slightly less slowly than the rate of ridge spreading.

20.3 The continental heat flux

Our understanding of the heat flux from continents is fundamentally different from the interpretation of ocean floor observations.

Continental rocks are much richer in radioactive elements and they extend much deeper, so that a substantial fraction of the heat flux from them is attributable to their own internal heat sources. This is illustrated by repeating with modern values a calculation by Strutt (1906), referred to in Section 4.2, to show that 23 km of 'standard granite', with $\rho = 2670\,\mathrm{kg\,m^{-3}}$ and radiogenic heat $1050\,\mathrm{W\,kg^{-1}}$, would provide the mean continental heat flux of $65\,\mathrm{mW\,m^{-2}}$. Also, much of the continental crust is so much older than any oceanic crust that a dynamic interpretation, with cooling of fresh igneous material, is relevant only to the limited areas of recent volcanism. That continents are close to being in a steady thermal state is indicated by the rather slight variation of heat flux with age of the continental basement (Fig. 20.4). The blocks in the figure show the scatter of large numbers of individual values, and the means with their formal (1 standard deviation) uncertainties are indicated by the central points for each age range. With the large number of data points and the care taken to ensure that the averages properly represent the different geological regions, we can see that there is only a slight age dependence, perhaps no more than could be attributed to the removal by erosion of surface layers with their radioactivity.

Old as they are, and steady as the heat flux from them appears to be, the continents cannot be regarded as static entities, passively floating with their lithospheric bases on fixed sections of the underlying mantle. The continents are moving, so that on the time scale of continental drift, 10^8 years, they slide, with their lithospheric plates, to different areas of the mantle. While the continental lithosphere may be close to an equilibrium state, the underlying asthenosphere is not. It is mechanically weak, allowing the lithospheric plates to slide across it and losing some heat to the continental areas but much more to the oceanic areas, where it progressively congeals on to the oceanic lithosphere. After 100 million years or so of cooling the oceanic lithosphere plunges back into the mantle at subduction zones, with renewal of the asthenosphere by material from below. Thus we must suppose that the deep thermal structures of the continents reflect to some extent the speeds of their motions and the times since they overlaid renewed asthenosphere. Compared with the ocean floor, the continents are thermal insulators so that the mantle beneath a slowly moving continent becomes hotter than in areas more recently exposed to ocean floor. Africa is of particular interest in this connection because it is very slow-moving. Evidence that the mantle under southern Africa and an adjacent area of ocean is hotter than in other areas is presented by Nyblade and Robinson (1994). They suggest that the extensive plateaux of eastern and southern Africa and elevation of the adjacent S. E. Atlantic ocean floor reflect high temperatures

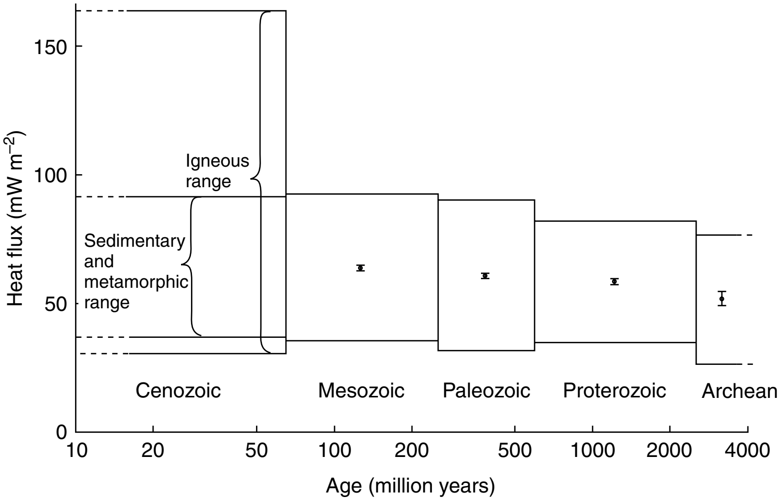


FIGURE 20.4 Continental heat flux. Ranges of values for terrains of different geological ages from Table 3 of Pollack *et al.* (1993).

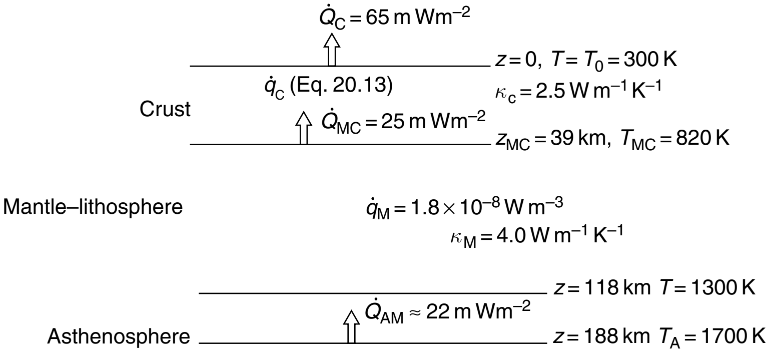

$\dot{Q}_C = 65$ m Wm^{-2}
$z = 0$, $T = T_0 = 300$ K
$\dot{q}_C$ (Eq. 20.13)
$\kappa_c = 2.5$ W m^{-1} K^{-1}
Crust
$\dot{Q}_{MC} = 25$ m Wm^{-2}
$z_{MC} = 39$ km, $T_{MC} = 820$ K
Mantle–lithosphere
$\dot{q}_M = 1.8 \times 10^{-8}$ W m^{-3}
$\kappa_M = 4.0$ W m^{-1} K^{-1}
$z = 118$ km $T = 1300$ K
$\dot{Q}_{AM} \approx 22$ m Wm^{-2}
Asthenosphere
$z = 188$ km $T_A = 1700$ K

FIGURE 20.5 A thermal model of continental lithosphere.

at depth over this broad area. Although the slow motion of Africa is a possible explanation, it is not the only one. Another is that Africa overlies a deep mantle plume that has not broken through to the surface, although the proximity of the hot spot marked by the volcanos Nyiragongo and Nyamarajira in south-east Zaire weakens the case for that alternative.

Although the continental heat flux is locally very variable, as Fig. 20.4 shows, the average is well constrained by observations, at least for areas with basement rocks of Mesozoic age or greater. We can use the average as the basis for a thermal model of the continents, in the approximation that they are old enough to be in a steady thermal state. With this assumption the surface heat flux is balanced by a combination of radiogenic heat in the crust and mantle and the heat flux into the base of the lithosphere from the asthenosphere (~118 km). Numerical details of the average model are shown in Fig. 20.5, which recognizes that the lithosphere is identified as a cooled and hardened layer, but that heat flows into it from some depth in the asthenosphere, where the temperature is assumed to be adiabatically related to the deeper mantle. This point is pursued in Section 20.4.

Radioactivity is strongly concentrated in the crust and is also subject to upward concentration within the crust. Uniform crustal radioactivity equal to the concentrations in surface layers would give more than the observed heat flux. Both the heat flux, Q_0, and radiogenic heating of surface rocks, q_0, are very variable, but they are observed to be correlated and the correlation leads to a simple model for the depth distribution of heat sources. Lachenbruch (1970) noted that a linear relationship

$$\dot{Q}_0 = A + B\dot{q}_0 \tag{20.12}$$

could be explained by assuming an exponential variation of radiogenic heat with depth,

$$\dot{q} = q_0 \exp(-z/z_0), \tag{20.13}$$

with the scale depth, z_0, approximately the same everywhere but q_0 variable. With this distribution the total crustal heat, Q_C, is the integral from the surface to the mantle–crust boundary at depth z_{MC},

$$\begin{aligned}\dot{Q}_C &= \dot{q}_0 \int_0^{z_{MC}} \exp(-z/z_0)\mathrm{d}z \\ &= \dot{q}_0 z_0[1 - \exp(-z_{MC}/z_0)] \approx \dot{q}_0 z_0, \end{aligned} \tag{20.14}$$

in which the simplification suffices because we require $z_{MC} \gg z_0$. Then $B = z_0$ and $A = \dot{Q}_{MC}$ is the heat flux from the mantle into the crust. Fitting of observations in different regions gives values of A in the range 20 to 30 m W m^{-2} and $B = 7$ to 10 km, so we here take averages, $\dot{Q}_{MC} = 25$ m W m^{-2} and $z_0 = 8$ km. We also assume an average crustal thickness, $z_{MC} = 39$ km, satisfying the condition $z_{MC} \gg z_0$. Subtracting $\dot{Q}_{MC}$ from the total heat flux, $\dot{Q}_0 = 65$ m W m^{-2}, we have $\dot{q}_0 z_0 = 40$ m W m^{-2}.

This simple model is convenient to use for some purposes, but its validity is disputed and in some obvious respects it is clearly unrealistic. It gives $(\dot{q}_0 z_0)/z_0 = \dot{q}_0 = 5 \times 10^{-6}$ W m^{-3}, exceeding the recognized standard value for granite, which is the most radioactive of the common rocks listed in Table 21.3. A common interpretation of the variations in $\dot{q}_0$ is that the same value applies to all fresh crust and that erosion removes surface layers to different depths, but this is contradicted by the very slight variation in average heat flux with age in Fig. 20.4. But, although Eq. (20.13) has serious shortcomings, simple alternatives fare no better. A uniform crust gives radiogenic heat exceeding the heat flux in large areas and a linear decrease with depth requires a cut-off at depths much less than z_{MC}. As long as Eq. (20.13) is applied cautiously, we can accept that it conveys some essential physics. Its usefulness is that it provides a consistent estimate of the mantle-to-crust heat flux $\dot{Q}_{\mathrm{MC}} = 25 \pm 5\ \mathrm{m\,W\,m^{-2}}$, in reasonable accord with the bottom of the range of surface heat flux in Fig. 20.4, $26\,\mathrm{m\,W\,m^{-2}}$.

With this simple model of crustal heat we can calculate the temperature profile. The heat flux through a surface at depth z is

$$\begin{aligned}\dot{Q} &= \kappa_{\mathrm{C}}\frac{\mathrm{d}T}{\mathrm{d}z} = \dot{Q}_{\mathrm{MC}} + \dot{q}_0 z_0 - \int_0^z \dot{q}\,\mathrm{d}z \\ &= \dot{Q}_{\mathrm{MC}} + \dot{q}_0 z_0 \exp(-z/z_0),\end{aligned} \qquad (20.15)$$

so that

$$\begin{aligned}T(z) - T_0 &= \frac{\dot{Q}_{\mathrm{MC}} z}{\kappa_{\mathrm{C}}} + \frac{\dot{q}_0 z_0}{\kappa_{\mathrm{C}}}\int_0^z \exp(-z/z_0)\mathrm{d}z \\ &= \frac{\dot{Q}_{\mathrm{MC}} z}{\kappa_{\mathrm{C}}} + \frac{\dot{q}_0 z^2}{\kappa_0}\exp(-z/z_0),\end{aligned} \qquad (20.16)$$

and at the mantle–crust boundary (with $z_{\mathrm{MC}} \gg z_0$)

$$T_{\mathrm{MC}} = T_0 + (\dot{Q}_{\mathrm{MC}} z_{\mathrm{MC}} + \dot{q}_0 z_0^2)/\kappa_{\mathrm{c}} = 820\ \mathrm{K}. \qquad (20.17)$$

Equation (20.17) is the starting point for a similar integration through the mantle component of the lithosphere. In this case we treat the lithospheric thickness as an unknown, to be estimated from an asthenospheric temperature, $T_{\mathrm{A}} = 1700\,\mathrm{K}$ at depth z_{A}, noting that this is really the temperature at some depth in the asthenosphere. This value is obtained by fitting the upper mantle to an adiabat anchored to the 660 km phase transition, with temperature increments at the other transitions as in Table G.3 (Appendix G). The mantle radiogenic heat is assumed to be uniform with a chondritic value, $\dot{q}_{\mathrm{M}} = 1.8 \times 10^{-8}\,\mathrm{W\,m^{-3}}$, at the upper mantle density. The radiogenic heat originating above z is $\dot{q}_{\mathrm{M}}(z - z_{\mathrm{MC}})$ and the heat flux through depth z is therefore

$$\dot{Q} = \dot{Q}_{\mathrm{MC}} - \dot{q}_{\mathrm{M}}(z - z_{\mathrm{MC}}) = \kappa_{\mathrm{M}}\mathrm{d}T/\mathrm{d}z. \qquad (20.18)$$

Integrating from z_{MC} to z_{A},

$$\begin{aligned}T_{\mathrm{A}} - T_{\mathrm{MC}} = (1/\kappa_{\mathrm{M}})[&Q_{\mathrm{MC}}(z_{\mathrm{A}} - z_{\mathrm{MC}}) \\ &- 1/2\,\dot{q}_{\mathrm{M}}(z_{\mathrm{A}} - z_{\mathrm{MC}})^2].\end{aligned} \qquad (20.19)$$

This gives a temperature gradient of 6 K/km in the sub-crustal lithosphere, starting from 820 K at the base of the crust (Eq. (20.17)). With the numerical values assumed, $(z_{\mathrm{A}} - z_{\mathrm{MC}}) = 149\,\mathrm{km}$, making $z_{\mathrm{A}} = 188\,\mathrm{km}$. This is the effective depth from which heat flows into the continental lithosphere and, as discussed in Section 20.4, is much greater than the mechanical thickness of the lithosphere. As mentioned, the heat generated in the mantle layer, $\dot{q}_{\mathrm{M}}(z_{\mathrm{A}} - z_{\mathrm{MC}}) \approx 2.7\,\mathrm{m\,W\,m^{-2}}$, is a small quantity, lost in the uncertainties.

The estimated flux of mantle heat into the continental crust, $25\,\mathrm{m\,W\,m^{-2}}$, is about a quarter of the average heat flux through the ocean floors and is a significant component of the energy budget of the mantle. Applying the estimate of continental area by Pollack *et al.* (1993), who included the submerged continental margins, 0.394 of the Earth's surface area or $2.0 \times 10^{14}\,\mathrm{m}^2$, this component of the heat flux is 5.0×10^{12} W. The difference between this and the total heat flux from the continents, $65\,\mathrm{m\,W\,m^{-2}}$, is attributed to crustal radioactivity. Adding a small contribution from the oceanic crust, the total crustal radiogenic heat is estimated to be 8.2×10^{12} W.

20.4 Lithospheric thickness

Although the lithosphere is identified as the surface layer that is cool enough to resist deformation, there is no sharp boundary separating it from the underlying asthenosphere. These layers do not differ in composition but only in temperature. Using the calculation in Section 20.3, we estimate the temperature gradient near to the base of the continental lithosphere as 6 K/km, which is almost ten times the melting point gradient and causes progressive softening with depth, which we represent as decreasing viscosity. For the thinner oceanic lithosphere the corresponding temperature gradient is about 12 K/km. But the ability of a material to resist deformation depends on the duration of the stress applied to it, so different lithospheric thicknesses are inferred from different observations. We can understand this in terms of the creep law represented by Eq. (10.27).

By assuming $n = 1$ in Eq. (10.27) we consider the linear (Newtonian) viscous regime, which is the simplest situation. Although we cannot be sure that this is valid at the base of the lithosphere, our immediate interest is in the effect of temperature and that can be examined without precluding a more complicated stress dependence. With this assumption Eq. (10.27) can be rewritten to make viscosity the subject:

$$\eta = \sigma/\dot{\varepsilon} = (\mu/B)\exp(gT_M/T). \qquad (20.20)$$

The strong temperature dependence of η arises from the fact that $g \gg 1$. Laboratory observations suggest $g \approx 27$, although the thermal history calculation in Section 23.4 requires a lower value for application to the mantle as a whole (see Fig. 23.1). We must consider a very wide range of viscosity, so it is convenient to take logarithms,

$$\ln \eta = \ln(\mu/B) + gT_M/T, \qquad (20.21)$$

from which we see that η increases by a factor 10 if gT_M/T increases by $\ln 10 = 2.3$. We are considering material approaching its melting point and so take $T_M/T = 1.2$ and, with the crude assumption that $g = 27$, $gT_M/T = 32.4$, to which an increment of 2.3 means 7.1%. If $T = 1300$ K, some distance above the asthenospheric temperature estimated in Section 2.3, then the required temperature change is 92 K. The effect of the temperature gradient is reduced by the gradient in T_M and, taking both together, we have, for the purpose of Eq. (20.21), an effective temperature gradient of 5.1 K/km. The pressure dependences of μ and B have little effect; although we have no secure information, the pressure dependence of g is probably slight and we neglect it. Then a 10-fold change in η occurs over a depth range 92 K/5.1 K/km = 18 km. Although very rough, this calculation suffices to show why the thickness of the mechanically identified lithosphere is ambiguous by tens of kilometres. However, relative thicknesses inferred from similar observations in different areas are not ambiguous on this account.

Now consider two quite different observations of the oceanic lithosphere. Seismic surface wave propagation, which imposes stresses with durations of tens of seconds, sees softening as decreasing elastic moduli. It sees the lithosphere as growing to a thickness of about 100 km with increasing age. Since these are observations of relative thickness by the same observations, the report by Zhang and Tanimoto (1991) that the thickness continues to grow in mature lithosphere presents a secure conclusion. On the other hand, the flexural response of the lithosphere to stresses with durations of millions of years, caused by superimposed loads such as chains of seamounts and volcanic islands, notably Hawaii, indicates thicknesses that may be only a third of the seismological thickness. Turcotte and Schubert (2002) estimate a thickness of 34 km for the lithosphere beneath the Hawaiian island chain from the flexural response to the load. The difference can be interpreted as the selection of boundaries marked by different isotherms and corresponding viscosities. We can use these observations to estimate the effective lithospheric thickness for flexure accompanying post-glacial rebound (Section 9.5). Using Eq. (20.21), we see that the ratio T_M/T at the effective base of the lithosphere is linearly dependent on the logarithm of the time scale of deformation. The time scale for seismic surface waves is tens to hundreds of seconds ($\tau_S \approx 10^{-6}$ year). The lithosphere under Hawaii responds to the superimposed load with a relaxation time that is not well observed, but must be less than the age of the youthful end of the island chain ($\tau_H < 10^7$ years) and is likely to be no more than the relaxation time for rebound (~5000 years). Thus, with subscripts S for surface wave data, H for the Hawaii observations and R for rebound, and assuming that the product (gT_M) in Eq. (20.21) is the same in all these situations, the temperatures at the effective bases of the lithosphere are related to the relaxation times by

$$\frac{1/T_{\mathrm{R}} - 1/T_{\mathrm{H}}}{1/T_{\mathrm{S}} - 1/T_{\mathrm{H}}} = \frac{\ln(\tau_{\mathrm{R}}/\tau_{\mathrm{H}})}{\ln(\tau_{\mathrm{S}}/\tau_{\mathrm{H}})} < 0.25, \qquad (20.22)$$

where the limit corresponds to $\tau_{\mathrm{H}} < 10^7$ years and could well be zero if the effective lithospheric basal temperature for rebound is the same as for the Hawaii situation.

To relate Eq. (20.22) to the lithospheric thicknesses, we note that τ_{S} and τ_{H} both refer to oceanic lithosphere, where the temperature gradient is steeper and the basal temperature correspondingly shallower than in the continental areas where rebound is observed. It is also possible that the lithosphere under Hawaii is volcanically heated over a broad enough area to influence the flexural thickness. Allowing this as an uncertainty and using the continental thermal profile calculation in Section 20.3, the flexural thickness of the lithosphere in areas of well-observed rebound is estimated to be 60 ± 15 km.

20.5 Climatic effects

The Earth's surface temperature varies with long-term climatic changes as well as the obvious diurnal and annual cycles and the variations propagate into the crust. In all cases the depth of penetration is very small compared with the dimensions of the Earth and the effect is described by the one-dimensional diffusion equation (Eq. (20.1)). The annual cycle penetrates only a few metres and is of little interest, but temperatures in the top few hundred metres of the continental crust reflect climatic variations extending back for hundreds to thousands of years. Pollack and Huang (2000) reviewed the studies directed to reconstructing past climates from borehole temperatures. The measurements are the same as those used to estimate the heat flow from the deep interior and the data analysis requires the separation of fluctuations from the steady gradient. There is no non-linear interaction between these effects and mathematically they can be treated as independent, so the subtraction is, in principle, simple, but disturbing effects such as variable rock properties introduce noise that limits what can be achieved.

If we consider a sinusoidal surface temperature variation, $T_0 \sin \omega t$, then we can represent its downward propagation with a combination of attenuation and phase lag,

$$T(t,z) = T_0 \exp(-\alpha z) \sin(\omega t - \beta z). \qquad (20.23)$$

By taking derivatives with respect to t and z and substituting in Eq. (20.1), we find $\alpha = \beta = (\omega/2\eta)^{1/2}$, that is

$$T(t,z) = T_0 \exp\left(-z\sqrt{\frac{\omega}{2\eta}}\right) \sin\left(\omega t - z\sqrt{\frac{\omega}{2\eta}}\right). \qquad (20.24)$$

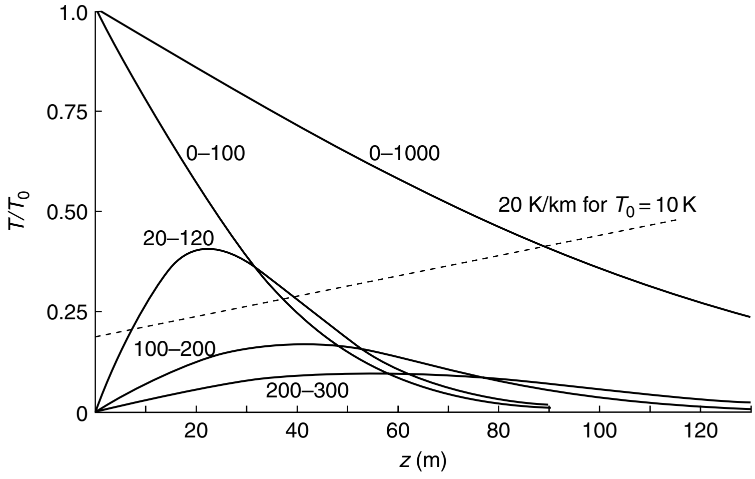


FIGURE 20.6 Penetration of the continental crust by surface temperature increments, T_0, lasting from 0 to 100, 0 to 1000, 20 to 120, 100 to 200 or 200 to 300 years ago.

This can be used directly for any regular oscillation that can be Fourier analysed, and, in principle, could be used for irregular variations, but a better method is to represent surface temperature changes by a sum (or integral) of a sequence of steps, each of which propagates downwards in the manner of Eq. (20.3). This is an example of an inverse problem for which the forward solution (calculation of a deep temperature profile from a known surface temperature variation) is straightforward and trial solutions can be repeatedly adjusted to find surface variations that best reproduce the observed profile. This is the trial-and-error method first applied by V. Cermak. Although more sophisticated inversion procedures are now used, they do not give identical results and there is no unambiguous 'best' method.

Strengths and limitations of the method are indicated by the hypothetical examples in Fig. 20.6, which shows temperature increments as functions of depth for five surface thermal impulses, all of magnitude T_0 and different durations. The curves for impulses starting either 100 or 1000 years ago and continuing to the present time are simply plots of the complementary error function, that is $(1 - \mathrm{erf}\, x)$, where erf x is given by Eq. (20.4). Note that they are scaled in depth by a factor $\sqrt{10}$ with respect to one another. For the purpose of calculation the 20 to 120-year plot assumes an impulse T_0 that began 120 years ago and continues to the present time but with a superimposed impulse $-T_0$, cancelling it 20 years ago. The 100 to 200-year and 200 to 300-year plots are calculated similarly. All these curves are normalized to the arbitrary impulse magnitude T_0. If $T_0 = 10\,\mathrm{K}$ the approximate magnitude of the steady gradient against which such effects must be observed is shown by the broken line.

As is apparent in Fig. 20.6, temperature variations that are brief compared with the times that have elapsed since they occurred are smeared out and become insignificant in a borehole profile. All of the curves can be re-scaled for times differing by any factor f from those plotted by re-scaling depth by a factor $\sqrt{f}$. This means that gradients change by $1/\sqrt{f}$, and since it is gradient variations that are sought, because they must be distinguished from the background gradient, resolution decreases with increasing age. In reviewing the results of numerous measurements, Pollack and Huang (2000) emphasized that temperature profiles from individual boreholes are subject to various disturbances and are not useful, but that systematic variations in numerous holes from a limited region reliably represent variations in surface temperature for that region.

The most obvious variation is a warming for the last 150 years, accelerating in the last 50 years, which is seen everywhere, most noticeably at high latitudes. An earlier cool period is consistent with a 'little ice age' from about 1500 to the mid nineteenth century, that probably had two separated maxima, although that would be difficult to see in the borehole data. The homogeneity of Greenland ice has allowed deep holes there to give paleotemperature estimates extending back as much as 25 000 years, to the last glacial maximum, in spite of complications arising from ice compression and glacial flow.

21

The global energy budget

21.1 Preamble

If the Earth's internal heat were not maintained by radioactivity, the present rate of heat loss, 44.2×10^{12} W, would cause cooling at an average rate of about 120 K per billion years. Even over 4.5 billion years, cooling at this rate would have had only a moderate effect on the temperature of the deep interior. Although radioactivity is the dominant continuing source of internal energy, it is only topping up the primordial heat generated by the Earth's accretion. It is slowing the cooling rate but is not the reason why the interior is hot. But radioactivity is decreasing slowly with time and does not balance the heat loss. The Earth is cooling at a rate determined by the difference between the surface heat flux and the radiogenic heat. Mantle convection and tectonics are slowing down as the Earth cools and the rate of change is determined by this unbalance.

A substantial source of internal energy is necessary also to maintain core convection, which drives the geomagnetic dynamo. The Earth has had a magnetic field for at least 3.5 billion years and probably for its entire life. The power requirement for dynamo action is probably less than the drain on core energy by thermal conduction, and if the core has some radioactivity then its effect is to compensate for the conductive heat loss. The necessity for this depends on the core conductivity, which is still poorly determined.

The first thing to note is that the gravitational energy released in the original accretion of the Earth dwarfs all other energy sources. This may be calculated by building up the Earth in layers, starting at the centre. When the radius has reached the value r, with mass m, the gravitational potential at the surface is $-Gm/r$ and when a further layer of mass $\mathrm{d}m = 4\pi\rho r^2\mathrm{d}r$ is added there is a loss of gravitational potential energy

$$\mathrm{d}E_{\mathrm{G}} = -Gm\,\mathrm{d}m/r = -4\pi Gm\rho r\,\mathrm{d}r, \tag{21.1}$$

where $m = 4\pi\int_0^r \rho r^2\mathrm{d}r$ and $G = 6.67 \times 10^{-11}$ $\mathrm{m^3kg^{-1}s^{-2}}$ is the gravitational constant. For a spherical mass M and radius R this integrates to

$$E_{\mathrm{G}} = -fGM^2/R, \tag{21.2}$$

as in the calculation of solar energy in Section 4.1, with $f = 3/5$ if the density is uniform. For a uniform body with the size and mass of the Earth this gives -224×10^{30} J, but with the central concentration of mass in the real Earth there is an additional energy release so that integration over the observed density profile gives $f = 0.6654$ and

$$E_{\mathrm{G}} = -249 \times 10^{30}\mathrm{J}\ (41.6 \times 10^6\ \mathrm{J/kg}). \tag{21.3}$$

This is fortuitously close to the value for a sphere with density inversely proportional to radius (Problem 21.1), for which $C/Ma^2 = 1/3$ (Problem1.1(c)), closely representing the central concentration of mass within the Earth. However, there is no formal relationship between moment of inertia and gravitational energy. These energies are written as negative because the zero of gravitational energy refers to infinite separation of material, making all values of E_{G} negative, but in most of what follows we are

Table 21.1 A comparison of global energies. All values are given in units of 10^{30} J. The first four entries give gravitational energy release with the resulting elastic strain energies subtracted. Strain energy is listed separately, so that the total gravitational energy release is the sum of the first five items. Tidal dissipation is included in the table but occurs mainly in the sea and does not influence the thermal state of the solid Earth

Accretion of a homogeneous mass	219.0
Core separation less strain energy	13.9
Inner core formation	0.09
Mantle differentiation	0.03
Elastic strain	15.8
Radiogenic heat in 4.5×10^9 years	7.6
Residual stored heat	13.3
Heat loss in 4.5×10^9 years	13.4
Present rotational energy	0.2
Tidal dissipation in 4.5×10^9 years	~ 1.1

interested in the positive gravitational energy released by aggregation or collapse of material. The magnitude of the number in Eq. (21.3) is emphasized by comparing it with the total release of radiogenic heat in the life of the Earth, $\sim 7.6 \times 10^{30}$ J (Table 21.1). If the energy in Eq. (21.3) were supposed to be retained by the Earth, it would correspond to an average temperature exceeding 37 000 K. Almost all of it was radiated away in the accretion process, but retention of a small fraction sufficed to ensure that the Earth started its life hot.

A consideration of gravitational energy also makes it difficult to avoid the inference that formation of the core accompanied accretion and was not delayed until the Earth was essentially complete. As pointed out in Section 5.4, the energy released by core separation from an initially homogeneous Earth would be 16×10^{30} J, sufficient to raise the temperature of the core by 6300 K. The gravitational instability of a homogeneous Earth, implied by these numbers, disallows any serious suggestion that it ever existed.

These calculations are necessarily numerical, with repeated iteration to the final version of any model, using equations of state for core and mantle materials to allow for their self-compression (Stacey and Stacey, 1999). An interesting by-product of the calculations is the energy of compression, which is stored irretrievably in the Earth as elastic strain energy. The total is 15×10^{30} J. In the core it averages 4.3×10^6 J kg^{-1}, and is comparable to the binding energy of iron (at zero pressure) or to the energy of a chemical explosive. In core processes involving redistribution of mass, as in the formation of the inner core, stored strain energy accounts for 12% of the gravitational energy release; this fraction produces neither heat nor dynamo power.

The principal continuing internal source of energy is radioactivity, which is slowly decaying. It yields about 28×10^{12} W at the present time but four times as much when the Earth was young and twice as much when averaged over the life of the Earth. Although it is an important contribution to the total energy budget, it is not an overwhelming one, as seen in the comparisons in Table 21.1. Although the Earth is hot inside because of the original accretion energy, two thirds of the current heat loss is attributable to radioactivity. A thermal history calculation (Chapter 23) is needed to tie these estimates together.

21.2 Radiogenic heat

The thermally important elements in the Earth are uranium, thorium and potassium, for which heat outputs are listed in Table 21.2. Their concentrations in the Earth are very uneven and many details remain obscure, but there are several considerations that allow us to develop a broad picture of their distributions. A comparison of the measured abundances in various geological materials is given in Table 21.3 with estimated average concentrations in the major components of the Earth. The thermal structure and history of the Earth depend on these concentrations, but, conversely, arguments about thermal history impose a constraint on the estimates of radiogenic heat. In particular, the thermal budget (Table 21.4) must balance.

A basic problem is to estimate the concentrations of radioactive elements in the mantle,

Table 21.2 Thermally important radioactive elements in the Earth

Isotope	Energy/atom[a] (MeV)	μW/kg of isotope	μW/kg of element	Estimated total Earth content (kg)	Total heat (10^{12} W)	Total heat 4.5×10^9y ago (10^{12} W)
^{238}U	47.7	95.0	94.35	12.86×10^{16}	12.21	24.5
^{235}U	43.9	562.0	4.05	0.0940×10^{16}	0.53	44.4
^{232}Th	40.5	26.6	26.6	47.9×10^{16}	12.74	15.9
^{40}K	0.71	30.0	0.003 50	7.77×10^{20}	2.72	33.0
				(Total K)	28.2	117.8

[a] These energies include all series decays to final daughter products. Average locally absorbed energies are considered; neutrino energies are ignored.

Table 21.3 Average radiogenic heat in geological materials. These numbers may be compared with the total heat flux per unit mass of the Earth, 7.4×10^{-12} Wkg^{-1}

		Concentration (parts per million by mass)				Heat production 10^{-12} W kg^{-1}
	Material	U	Th	K	K/U	
Igneous rocks	granites	4.6	18	33 000	7 000	1050
	alkali basalts	0.75	2.5	12 000	16 000	180
	tholeitic basalts	0.11	0.4	1500	13 600	27
	eclogites	0.035	0.15	500	14 000	9.2
	peridotites, dunites	0.006	0.02	100	17 000	1.5
Meteorites	carbonaceous chondrites	0.020	0.070	400	20 000	5.2_3
	ordinary chondrites	0.015	0.046	900	60 000	5.8_5
	iron meteorites	nil	nil	nil	–	$<3 \times 10^{-4}$
Moon	Apollo samples	0.23	0.85	590	2 500	47
Global averages	crust (2.8×10^{22} kg)	1.2	4.5	15 500	13 000	293
	mantle (4.0×10^{24} kg)	0.025	0.087	70	2800	5.1
	core	nil	nil	29	–	0.1
	whole Earth	0.022	0.081	118	5400	4.7

where they cannot be directly observed. As evident from the numbers in Table 21.3, the ratio of thorium to uranium is quite similar in a wide range of rock types as well as meteorites. Although lower Th/U ratios occur in mid-ocean ridge basalts (MORB), it is reasonable to suppose that the global balance of these elements has preserved the meteorite ratio. Here the overall mantle ratio is assumed to be 3.7. Estimates of the potassium/uranium ratio are less secure. Although both of these elements are strongly concentrated in the crust, the evidence of argon outgassing, considered below, indicates that potassium is more strongly depleted in the mantle than are uranium and thorium. We cannot appeal to meteorite data for an estimate of the global potassium content, because, being a volatile element, it is certainly depleted in the Earth relative to the meteorites. Even the meteorites show substantial variations, as do all of the

Table 21.4 The heat budget (all values in units of 10^{12} W)

Income			
Crustal radioactivity	8.2		
Mantle radioactivity	20.0		
Core radioactivity	0.2		
Latent heat and gravitational energy released by core evolution	1.0		
Gravitational energy of mantle differentiation	0.1		
Gravitational energy released by thermal contraction	3.1		
Tidal dissipation	0.1		
TOTAL		32.7	
Expenditure			
Crustal heat loss	8.2		
Mantle heat loss	32.5		
Core heat loss	3.5		
TOTAL		44.2	
Net Loss of Heat			11.5

Solar System bodies that have been sampled (Fig. 21.1).

Fisher (1975) drew attention to the fact that argon and helium trapped in glassy (rapidly cooled) submarine basalts indicated that the K/U ratio of the mantle, from which they were derived, may be as low as 1500. More recent work, involving simultaneous measurements of ^{4}He, ^{40}Ar and of the non-radiogenic isotopes ^{3}He and ^{36}Ar, which evidently remain in the Earth from the original accretion, has given very variable results which have not been untangled from the problem of contamination by the rare gases, especially ^{3}He, in marine sediment introduced by the infall of cosmic dust. There is no compelling reason to adopt a particular value for the mantle average K/U ratio. Values in the range of 1500 to 6000 appear plausible and 2800 is assumed in Tables 21.2 to 21.4.

An extreme lower bound on total terrestrial potassium (4.7×10^{20} kg) is imposed by the ^{40}Ar in the atmosphere. This is only slightly greater than the estimate of crustal potassium in Table 21.3. A simple outgassing history is compatible with a total potassium content of the Earth slightly less than twice this value (Section 5.2). We have no comparable check on the outgassing of ^{4}He from the Earth, as helium rapidly escapes from the atmosphere to space.

It has been conventional to assume negligible radioactivity in the core, essentially because it is not found in meteoritic iron, but there is a long history of suggestions that the core contains potassium, associated with sulphur, which is almost certainly present in both the outer and inner cores. As we mention in Section 2.8, the chemical argument for a substantial potassium content in the core is not conclusive. Also, we follow the conclusion of Wheeler *et al.* (2006) that the uranium content of the core is negligible. The physical case for core radioactivity is based on the need for a heat source to compensate for the conductive heat loss and that requires an estimate of thermal conductivity (Section 19.6). This is not known with sufficient accuracy to be clear whether radiogenic heat is needed, but current estimates favour a small contribution and 0.2 terawatt of heat from ^{40}K is included in Tables 21.3 to 21.5. This is the $A = 0.2$ model in Section 21.4.

Uranium and thorium are not readily separated by volatility or normal geochemical processes and appear in meteorites, and by implication in the Earth, in an almost constant ratio. This is evidence that they are not fractionated relative to other non-volatiles and so have approximately meteoritic abundances in the Earth. With the radiogenic heat that this implies and the total heat required by thermal history calculations (Chapter 23), we have an estimate of potassium abundance independent of the evidence from ^{40}Ar/^{4}He ratios in mantle-derived igneous rocks. The numbers in Tables 21.3 and 21.4 satisfy these conditions but give rather less potassium than often suggested. We consider what the data in Fig. 21.1 indicate. On this graph, materials having equal concentrations of uranium lie on a line of gradient unity. The ordinary and carbonaceous chondrites are not far from such a line, encouraging the view that they have similar uranium abundances. But the ordinary chondrites are

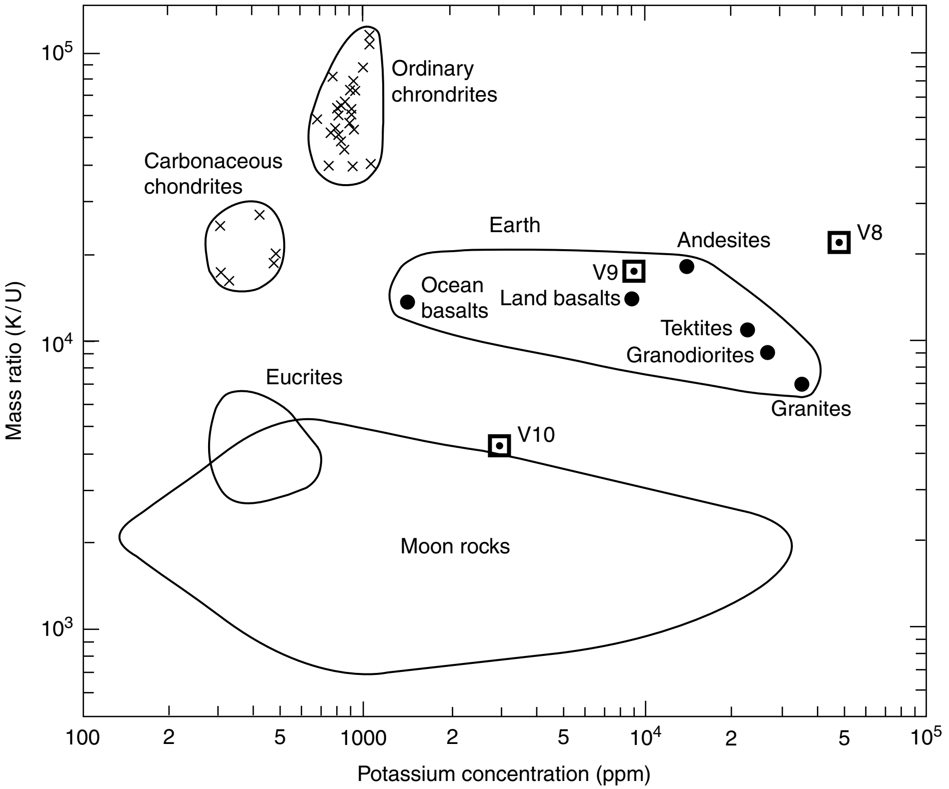


FIGURE 21.1 Mass ratios, K/U, as a function of K concentrations, showing the separate groupings of crustal rocks, ordinary and carbonaceous chondrites, eucrites (a class of achondrite), Moon rocks and three samples from Venera probes 8, 9 and 10. Most of the data are from Eldridge *et al.* (1974). Values for the Venus samples are from Keldysh (1977).

systematically richer in potassium, so there was evidently a processing of potassium that did not affect uranium. Variations in potassium abundance are commonly attributed to its volatility, but this does not readily explain its enrichment in ordinary chondrites, relative to the volatile-rich carbonaceous chondrites. The data for terrestrial samples, Moon rocks and the available Venus analyses are well removed from the trend of the chondrite data in Fig. 21.1, but they are for crustal materials, in which all of the radioactive elements are concentrated. They have been subjected to much more severe processing than the chondrites and, in view of the wide disparity between potassium concentrations in the different types of chondrite, we conclude that none of these data provide evidence of the total potassium content of the Earth. Evidence of its outgassing and the ^{40}Ar content of the atmosphere, with the ^{40}Ar/^{4}He ratios in glassy submarine basalts, provide the clearest information that we have.

21.3 Thermal contraction, gravitational energy and the heat capacity

The continued gradual cooling of the Earth is accompanied by thermal contraction and consequent release of gravitational energy. In thermal history calculations this is allowed for by an adjustment to the heat capacity; the gravitational energy is added to the heat lost per degree of cooling. The magnitude of the effect is the subject of this section. Although for the purpose of thermal history it is accounted for by the heat capacity adjustment, we need to note that, in balancing the energy budget, the gravitational energy increases the difference between the heat loss and the radiogenic heat. Geochemical estimates of the mantle content of thermally important elements (e.g. McDonough and Sun, 1995) give about 20×10^{12} W at the present time and thermal history calculations (Section 23.4) indicate that this is reasonable, so we adopt it for our thermal model. However, it could be in error by 4×10^{12} W and our estimate of the K/U ratio differs from that by McDonough and Sun, so this uncertainty carries through all our calculations.

The gravitational energy released by the contraction of any material element is readily calculated if it is assumed that the densities of all other elements are unaffected, even though their positions may change. However, this is not what happens. When the Earth contracts by thermal shrinkage of any component, the pressure is increased throughout, so that the density increases everywhere, adding to the shrinkage and to the consequent gravitational energy release. It is necessary to consider the Earth as a whole and to allow for compression as well as thermal shrinkage. A perturbation calculation appears possible in principle, but the obvious, direct way of doing this is to compare two independently calculated, self-compressed models with specified pressure–density relationships for each of the components. Then it is found that shrinkage of the mantle causes compression and therefore gravitational energy release in the core as well as in the mantle itself, and vice versa. Calculations of this kind were reported by Stacey and Stacey (1999) in a sufficiently general way to allow rescaling to any assumed contraction of the core or mantle, and the results are applied here.

We seek a relationship between gravitational energy release and the net energy loss, that is, the ratio of the third and second terms in the heat balance equation:

$$\text{radiogenic heat} + \text{net cooling} + \text{gravitational energy loss} - \text{compressional energy} = \text{surface heat flux}. \tag{21.4}$$

For this purpose mantle and core cooling must be treated independently because their cooling rates are not the same and the ratio of gravitational energy loss to cooling is bigger in the case of the core. The mantle case is simpler, being a straightforward application of the Stacey and Stacey (1999), results with only a minor influence of the core. We refer to the effective heat capacity of the mantle as the loss of heat required for the potential temperature of the lower mantle to fall by one degree, and this differs little from the potential temperature of most of the upper mantle. The point of this definition is that if the temperature gradient is everywhere adiabatic, and remains so with cooling, the decrease in temperature is proportional to the absolute temperature. Since the potential temperature, $T_\mathrm{p} = 1700\,\mathrm{K}$, is the temperature of adiabatic extrapolation to zero pressure, material at any higher pressure, being at a higher temperature, $T > T_\mathrm{p}$, contributes more to the heat loss than if uniform cooling is assumed. Using the thermal model in Appendix G, the effective heat capacity of the mantle is, without considering thermal contraction,

$$\begin{aligned}\text{Mantle}: m\langle C_P\rangle &= \int \rho C_P (T/T_\mathrm{p})\mathrm{d}V \\ &= 6.19 \times 10^{27}\ \mathrm{J\,K^{-1}}.\end{aligned} \tag{21.5}$$

Defined in this way, $\langle C_P\rangle = 1537\,\mathrm{J\,K^{-1}kg^{-1}}$, compared with $C_P \approx 1200\,\mathrm{J\,K^{-1}kg^{-1}}$, as calculated by Eq. (19.4) and listed in Appendix G. A similar calculation gives the corresponding thermal contraction

$$\begin{aligned}\text{Mantle}: V\langle \alpha\rangle &= \int \alpha (T/T_\mathrm{p})\mathrm{d}V \\ &= 2.12 \times 10^{16}\ \mathrm{m^3\,K^{-1}}.\end{aligned} \tag{21.6}$$

This corresponds to an effective expansion coefficient $\langle\alpha\rangle = 23.4 \times 10^{-6}\ \mathrm{K^{-1}}$, being the average contraction for a one degree fall in potential temperature.

From the results of the Stacey and Stacey (1999) calculations we see that a uniform 1% contraction of the mantle would release 5.90×10^{29} J of gravitational energy, but that 0.75×10^{29} J of this would become elastic strain energy, leaving a net release of 5.15×10^{29} J. With the value of $\langle\alpha\rangle$ calculated above, a 1% contraction corresponds to a fall in potential temperature of 427 K, so we have a gravitational contribution of 5.15×10^{29} J/427 K $= 1.21 \times 10^{27}$ J K^{-1} to the effective heat capacity, which, added to Eq. (21.5), gives a total effective mantle heat capacity

$$\phi_{\mathrm{m}} = 7.40 \times 10^{27}\ \mathrm{J\,K^{-1}}. \tag{21.7}$$

This is used in the thermal history calculations in Chapter 23. The numbers show that, of the net heat loss by the mantle, approximately 20% is gravitational energy, although this slightly overestimates the effect because α decreases with depth and the deep mantle is weighted by the factor T/T_{P} in the gravity calculation.

In considering the cooling of the core, the growth of the inner core introduces a complication, but we treat it as adding three independently calculable terms to the effective heat capacity, attributed to latent heat of solidification, gravitational energy of contraction due to solidification and the gravitational energy released by compositional separation of the light outer core solute. Section 21.4 examines the radial distribution of heat sources, but here we estimate their global effects. We use the core–mantle boundary temperature, T_{CMB}, as the reference, with the higher internal temperatures adiabatically related to it, as for T_{P} in the case of the mantle. We define the effective core heat capacity as the net heat loss per degree fall in T_{CMB}.

For the general cooling and contraction, without consideration of inner core development, we have expressions analogous to Eqs. (21.5) and (21.6):

$$\text{Core}: m\langle C_P \rangle = 1.80 \times 10^{27}\,\mathrm{J\,K^{-1}}, \tag{21.8}$$

$$V\langle \alpha \rangle = 2.79 \times 10^{15}\ \mathrm{m^3K^{-1}}. \tag{21.9}$$

Equation (21.9) corresponds to $\langle \alpha \rangle = 15.8 \times 10^{-6}$ K^{-1}. By the Stacey and Stacey (1999) calculations a 1% contraction of the core would release 5.36×10^{29} J, less 0.73×10^{29} J of strain energy, leaving 4.63×10^{29} J, and with the estimated $\langle \alpha \rangle$ this would require T_{CMB} to fall by $(0.01/15.8 \times 10^{-6})$ K $= 634$ K. Thus the gravitational contribution to the heat capacity is 4.63×10^{29} J/634 K $= 7.31 \times 10^{26}$ J K^{-1}. This is added to Eq. (21.8) to give an effective core heat capacity

$$\phi_{\mathrm{c}} = 2.53 \times 10^{27}\ \mathrm{J\,K^{-1}}, \tag{21.10}$$

with no allowance for inner core development. This is 40% greater than the heat capacity estimated with neglect of gravitational energy.

The three additional energy contributions mentioned above can be converted to notional components of the heat capacity by dividing them by the temperature change, expressed as the fall in T_{CMB} required for inner core formation. First we estimate the magnitudes of the energies. The latent heat of inner core formation, L, is calculated from the entropy by Eq. (19.53),

$$L = T_{\mathrm{M}}\Delta S = T_{\mathrm{M}}(nR\ln 2 + \alpha K_T \Delta V), \tag{21.11}$$

where $n = M_{\mathrm{IC}}/m$ is the number of moles of inner core material, with total mass M_{IC} and mean atomic weight $m = 50.16$ by Table 2.1. It is convenient that the product $\langle \alpha K_T \rangle$ varies rather little and we take an average from Appendices F and G:

$$\begin{aligned}\langle \alpha K_T \rangle &= \langle \alpha K_S/(1 + \gamma\alpha T) \rangle \\ &= 12.16 \times 10^6\ \mathrm{Pa\,K^{-1}}.\end{aligned} \tag{21.12}$$

As in Section 19.4, we take the density change on freezing of constant composition to be 200 kg m^{-3} (1.55% of the average inner core density), making $\Delta V = 1.18 \times 10^{17}$ m^3, and, with an average boundary temperature during solidification of 5050 K, these numbers give

$$L = 6.36 \times 10^{28}\ \mathrm{J}. \tag{21.13}$$

The gravitational energy released by the contraction caused by freezing was calculated by Stacey and Stacey (1999) for an assumed density increment of 140 kg m^{-3} to be 4.123×10^{28} J, less 0.515×10^{28} J of strain energy. Revising these values to a density increment of 200 kg m^{-3} we obtain the net gravitational energy release by solidification, 5.15×10^{28} J. This is the global energy release, some of which occurs in the mantle. For the purpose of calculating the energy balance of the core and the contribution to dynamo power (Section 22.7), we require only the fraction of this energy that is released in the core, and write this as

$$E_{\mathrm{GS}} = 3.14 \times 10^{28}\ \mathrm{J}. \tag{21.14}$$

A simple calculation of the energy released by compositional separation is given in Section 22.6, but we can use the results of the Stacey and Stacey (1999) calculation, with the advantage that the effects of global contraction and strain energy increase are included. The compositional component of the density contrast between the inner and outer cores is taken to be 620 kg m^{-3}, being a total density difference of 820 kg m^{-3}, as estimated by Masters and Gubbins (2003), less 200 kg m^{-3} for solidification. Renormalizing the Stacey and Stacey result to this density contrast, the global energy release minus strain energy is 4.99×10^{28} J, of which the fraction released in the core is

$$E_{\mathrm{GC}} = 4.79 \times 10^{28}\ \mathrm{J}. \tag{21.15}$$

We relate the energy released by inner core formation to its contribution to the global effective heat capacity by calculating the required cooling. This is the difference between the adiabatic and melting point gradients over the pressure difference between the centre and the present inner core boundary, but with an adjustment for the increase in pressure caused by the contraction. The principle is illustrated in Fig. 21.2. With the density increments used above there are increments in central pressure of 1.82 GPa and 4.41 GPa due to solidification and compositional separation, giving 6.23 GPa together. There are also contributions of 5.4×10^{-3} GPa K^{-1} by general core cooling (referenced to T_{CMB}) and 4.1×10^{-3} GPa K^{-1} per degree change in T_{P}, arising from thermal contraction of the mantle. These effects are all subtracted from the present pressure difference between the centre and the inner core boundary (35.05 GPa by PREM) to obtain the pressure change at the boundary as the inner core grows,

$$\begin{aligned}\delta P(\mathrm{GPa}) = (35.05 - 6.23) + 5.4 \\ \times 10^{-3}\Delta T_{\mathrm{CMB}}(\mathrm{K}) + 4.1 \\ \times 10^{-3}\Delta T_{\mathrm{P}}(\mathrm{K}).\end{aligned} \tag{21.16}$$

The cooling is calculated from the difference between the melting point gradient by Eq. (19.50) and the adiabatic gradient by Eq. (19.19) over the pressure range δP given by Eq. (21.16), with $T = T_{\mathrm{M}}$,

$$\begin{aligned}\Delta T_{\mathrm{ICB}} &= \left[\frac{\mathrm{d}T_{\mathrm{M}}}{\mathrm{d}P} - \left(\frac{\partial T}{\partial P}\right)_S\right]\delta P \\ &= \left[\frac{2\gamma T(1+\gamma\alpha T)}{K_S(1+2\gamma\alpha T)} - \frac{\gamma T}{K_S}\right]\delta P \\ &= \frac{\gamma T \delta P}{K_S(1+2\gamma\alpha T)},\end{aligned} \tag{21.17}$$

with adiabatic extrapolation giving

$$\Delta T_{\mathrm{CMB}}/\Delta T_{\mathrm{ICB}} = T_{\mathrm{CMB}}/T_{\mathrm{ICB}} = 0.787. \tag{21.18}$$

Solution of these equations requires an assumption about the mantle cooling (ΔT_{P} in Eq. 21.16) that occurs during inner core formation, but its effect is small enough that no serious uncertainty arises. We take $\Delta T_{\mathrm{P}} = 300$ K. Then Eqs. (21.16) to (21.18) give

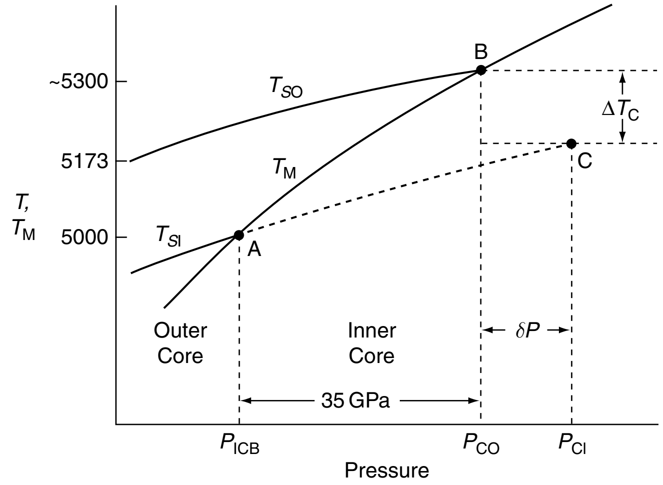


FIGURE 21.2 Temperature–pressure relationship for the inner core. The outer core adiabat, T_{S1}, intersects the melting curve, T_{M}, at A and is extrapolated to the centre of the Earth at C, pressure P_{C1}. The hotter adiabat at commencement of inner core growth, T_{S0}, meets T_{M} at central pressure P_{C0} (at point B), which is less than P_{C1} because the whole Earth was then hotter and more dilated. Growth of the inner core to its present size requires cooling ΔT_{C} at the centre of the Earth and $0.723\,\Delta T_{\mathrm{C}}$ at the core–mantle boundary.

$$\Delta T_{\mathrm{CMB}} = 98.4\ \mathrm{K}. \quad (21.19)$$

Now we add to the effective heat capacity by Eq. (21.10) the notional contributions by inner core formation:

$$\varphi_{\mathrm{cTotal}} = \varphi_{\mathrm{c}} + (L + E_{\mathrm{GS}} + E_{\mathrm{GC}})/\Delta T_{\mathrm{CMB}} \\ = 4.21 \times 10^{27}\ \mathrm{J\,K^{-1}}. \quad (21.20)$$

This is the effective core heat capacity, averaged over the duration of inner core formation. The value in Eq. (21.10) must be used for the period, if there was such early in the life of the Earth, when there was no inner core. Sections 21.4 and 22.7 consider the radial variation of heat sources. The values estimated here are global totals, with separate identification of energy released in the core that is needed in calculating dynamo power and discounting the energy of elastic compression.

We must note that the numbers in this section present a simplification of the effective core heat capacity, which increases with time as the inner core grows. The value in Eq. (21.20) is an average over the period of development of the inner core and the fractional contribution of $(L + E_{\mathrm{GS}} + E_{\mathrm{GC}})$ to the total has increased with time. These terms are contributed by the growth in inner core volume, which is not linear in core temperature. This problem is referred to in Section 22.7 and discussed in more detail in Section 23.5.

21.4 Energy balance of the core

The energy estimates in Section 21.3 are now reexamined to determine their distribution within the core. A basic question is: how fast is it cooling? We know that it must have been losing heat fast enough to maintain dynamo action for several billion years, and this requires three-dimensional stirring of the outer core that keeps the temperature gradient very close to adiabatic. There is, therefore, a conducted heat flux at all levels, and this is a base load on core energy that must be supplied, whatever combination of thermal and compositional driving forces is responsible for the convection. The conducted heat contributes nothing to dynamo action, which requires additional energy, thermal, compositional, gravitational, or some combination of them to drive convection. The compositional contribution is particularly important, being mechanical energy of gravitational origin that is 100% efficient in driving the dynamo (discounting a small loss by diffusive mixing). Thermal convection operates with a limited thermodynamic efficiency (12% to 25%, depending on the distribution of the heat source) and exhausts, as heat to the core–mantle boundary, much more energy than it produces as convective power. With a strong compositional effect the heat flux into the mantle may be less than the conducted heat at the top of the core. In this situation compositional convection carries back down some of the conducted heat, a process that we refer to as refrigerator action. It is needed only over a limited depth range at the top of the core and therefore a limited temperature range, making it a very efficient process in the thermodynamic sense of requiring little mechanical energy for a large heat transfer.

We identify seven sources of energy in the core. Table 21.5 gives their integrated contributions over the lifetime of the inner core, τ, which may be less than the age of the Earth, τ_{E}. By listing the total energies in this way we postpone, to Sections 22.7 and 23.5, consideration of variations in the rates at which they are dissipated. Here we consider average rates over time τ, obtained by dividing the energies by τ. The first five entries in the Q column of Table 21.5 are discussed in Section 21.3. Precessional dissipation is considered in Section 7.5 and re-examined in Section 24.7 as a possibly significant contributor to the dynamo early in the life of the Earth. The entry in the table, $0.20\tau/\tau_{\mathrm{E}}$, corresponds to a mean dissipation rate of 0.014 terawatt, a little more than twice our estimate of the present rate but only 20% of the very early rate. It is only a minor contribution.

The least certain of the entries in Table 21.5 is radioactivity. We follow the chemical argument in Section 2.8 by assuming that any radiogenic heat is due to ^{40}K, but allow the possibility that it is zero. In the table it is represented by the parameter A, which is the present heat output

Table 21.5 Summary of components of the core energy balance during the lifetime τ of the inner core. Q gives energy sources and E the corresponding convective energies, which are related to Q by thermodynamic efficiencies. τ_E is the age of the Earth and A is the present rate of radiogenic heating by ^{40}K in terawatts. For a discussion of refrigerator action see Section 22.8

Sources		Q (10^{28} J)	Efficiency	E (10^{28} J)
Heat capacity	Cooling	17.75	0.139	2.47
	Contraction	2.52	0.135	0.34
Freezing	Latent heat	6.36	0.252	1.60
	Contraction	2.76	0.190	0.52
Compositional separation		4.79	1.00	4.79
Precession		$0.20\ \tau/\tau_E$	1.00	$0.20\ \tau/\tau_E$
Radioactivity		$5.69A[\exp(2.5\ \tau/\tau_E) - 1]$	0.127	$0.723A[\exp(2.5\ \tau/\tau_E) - 1]$
Losses				
Conduction		$54.0\ \tau/\tau_E$	−0.119	$-6.43\ \tau/\tau_E$
Refrigerator action		$-Q_R$	f	$-f\,Q_R$

in terawatts. The expression in the table is the integral over time τ, with a numerical factor to give the integrated heat in units of 10^{28} J, as for other entries in the table. We consider two models, with $A = 0.2$ terawatt and $A = 0$.

The major item of heat loss listed in Table 21.5, $54.0\tau/\tau_E$, is the conducted heat at the top of the core, being 3.79×10^{12} W for time τ, assuming a conductivity of $28.3\,\mathrm{W\,m^{-1}K^{-1}}$ (Section 19.6) and an adiabatic temperature gradient (Eqs. (19.55) and (19.56)). Arguments about the energy balance of the core, and the need to consider radioactivity, are critically dependent on this heat loss, and therefore on the conductivity. Most discussions have assumed a higher value. The nature of the problem is seen by noting that if the inner core has existed for most of the life of the Earth ($\tau \approx \tau_E$), then the sources in the Q column of Table 21.5 balance the heat loss only with a radiogenic contribution, refrigerator action, or both. Since refrigeration consumes convective power that would otherwise be available for the dynamo (the $-fQ_R$ entry in the E column of the table), its availability is limited. To determine the efficiency factor f with which it operates we need to know the radial variation of the core heat flux.

The heat flux within the core is plotted in Fig. 21.3 for the two models considered, with and without radiogenic heat. Conducted heat diminishes strongly with depth mainly because the adiabatic temperature gradient decreases (Eq. 19.55). The total heat flux through any radius is the total heat originating within it and varies less with radius than the conducted heat, both because of the latent heat of inner core solidification and because the deep parts cool faster in maintaining the adiabatic gradient. In the figure we see that the total heat comfortably exceeds the conducted heat deep in the core but that for both models the conducted heat is greater in the outer part. Convected heat, shown as the difference, is negative at the top of the core (shaded for the $A = 0.2$ model in Fig. 21.3), where the adiabatic temperature gradient must be maintained by reversed convection, that is refrigerator action, the principle of which is discussed in Section 22.8. The refrigerator power required is the product of heat carried down and an efficiency, f, determined by the temperature range over which it operates. f is the Carnot efficiency with which the same heat would generate power if convected upwards. For the $A = 0.2$ model $f = 0.043$. For the $A = 0$ model not only must more heat be transferred downwards but the depth and temperature ranges are greater and $f = 0.091$. With these values of f we can examine the models to see whether the refrigerator action leaves enough convective energy for a viable dynamo.

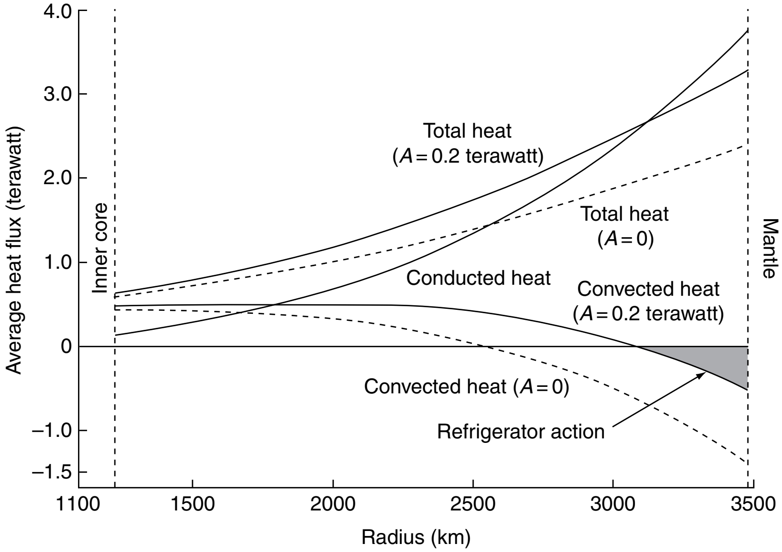


FIGURE 21.3 Radial variation of the average core heat flux over the life of the Earth for two models, both of which assume that the inner core began to form very early. Conducted heat, calculated assuming a thermal conductivity which varies from 28.3 W m^{-1}K^{-1} at the top to 29.3 W m^{-1}K^{-1} at the bottom of the outer core, exceeds the plotted total heat flux in the outermost part of the core for both models. The difference is plotted as convected heat, the negative range indicating how much heat is carried down by refrigerator action, driven by compositional convection. The model with 0.2 terawatt of present radiogenic heat from ^{40}K (the $A = 0.2$ model) is shown by solid lines and the non-radioactive model by broken lines.

The sources and losses in the Q column of Table 21.5 must balance. With A in terawatts and other numerical values adjusted so that all terms are expressed in units of 10^{28} J, as in the table,

$$34.14 + 5.69A[\exp(2.5\ \tau/\tau_E) - 1] = 53.8\tau/\tau_E - Q_R. \tag{21.21}$$

Now we need also the entries in the E column of the table. These are convective energies related to corresponding entries in the Q column by thermodynamic efficiencies discussed in Section 22.7. The sum of all the E entries (subtracting the negative terms) is the energy, E_D, available for the dynamo. But the dynamo itself generates heat by ohmic dissipation; by assuming it to be uniformly distributed through the core we assign an efficiency of 0.122 to its generation of further dynamo power. This cannot be counted as a heat source but it means that we equate the available energy not to E_D but to $(1 - 0.122)\ E_D$. Thus, in units of 10^{28} J as before,

$$\begin{aligned} 0.878E_D = 9.72 + 0.723A[\exp(2.5\tau/\tau_E) - 1] \\ - 6.23\tau/\tau_E - fQ_R. \end{aligned} \tag{21.22}$$

Substituting for Q_R by Eq. (21.21) we obtain an expression for dynamo power in terms of A, τ and f. For the two cases considered, ($A = 0$, $f = 0.091$) and ($A = 0.2, f = 0.043$),

$$A = 0 : E_D = 14.61 - 12.67\tau/\tau_E, \tag{21.23}$$

$$\begin{aligned} A = 0.2 : E_D = 12.66 - 9.73\tau/\tau_E \\ +0.2204[\exp(2.5\tau/\tau_E) - 1]. \end{aligned} \tag{21.24}$$

Dividing E_D by τ, we have the average dynamo power for the lifetime, τ, of the inner core as a function of τ for each of the models, as plotted in Fig. 21.4. The broken line in the figure shows the estimated dynamo power requirement from Section 24.7, 0.3 terawatt. Remembering that we have deferred to Sections 22.7 and 23.5 the variations of core heat flux and dynamo power with time, the figure shows the average power over the inner core lifetime and not a time variation of the power. The essential conclusion from this figure is that the assumed core conductivity, $\kappa = 28.3$ W m^{-1}K^{-1} at the top of the core and varying little with depth, marks a boundary between alternative core models, with and without radioactivity. With a lower conductivity there would be no reason to consider radioactivity, but a higher value would make it difficult to avoid. The uncertainty in κ is the central doubt around which the current debate about core physics revolves, as reviewed by Nimmo *et al.* (2004).

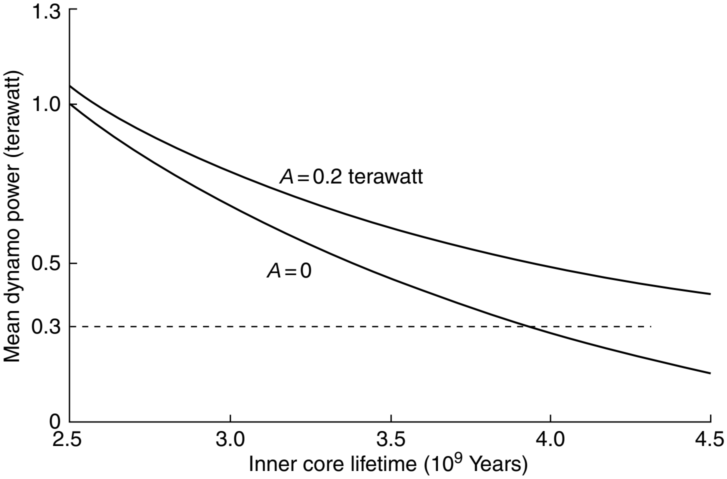


FIGURE 21.4 Average dynamo power over the inner core lifetime, *t*, as a function of *t* for models with 0.2 terawatt of radiogenic heat from ^{40}K at the present time and no radiogenic heat ($A = 0$). The broken line indicates the power requirement for a geomagnetic field of the present strength.

21.5 Minor components of the energy budget

By ‘minor components’ we mean contributions that may be geophysically interesting and are included in Table 21.4, but are smaller than the uncertainties in the major components and are not discussed elsewhere in this chapter. We consider two: tidal friction and mantle differentiation. Others that we can think of, such as absorption of cosmic neutrinos, are certainly insignificant.

The tidal energy dissipation, calculated from the slowing rotation in Section 8.3, with a small allowance for the increasing orbital energy of the Moon, is 3.7×10^{12} W. Most of this is due to marine tides and has no impact on the solid Earth. The internal dissipation can be estimated from the elastic strain energy of the solid tide, of which a fraction, $2\pi/Q_S$, is lost per tidal cycle by anelastic effects. We adopt the value of $Q_S = 280$ for shear deformation of the mantle at tidal frequencies obtained by Ray *et al.* (2001), who combined satellite observations of the total tidal deformation of the Earth, as deduced from its gravitational effect, with altimeter data on the deformation of the ocean surface. Although measuring a small effect as a difference between much larger ones, they claimed 25% accuracy. Their tidal Q_S value appears high compared with what might be expected from extrapolation from seismic frequencies, but it is the only direct observation. The tidal strain of the solid Earth, estimated from the Love number *h* (Section 10.5), is $\varepsilon = 5 \times 10^{-8}$ so the strain energy is $\frac{1}{2}\mu\varepsilon^2 V = 2 \times 10^{17}$ J, with rigidity modulus, $\mu = 176$ GPa, averaged over the volume, *V*, of the mantle. With $Q_S = 280$, the fraction of this energy that is dissipated each tidal cycle (12.4 hours or 44 700 s) is 2.25%, giving a rate of 1.0×10^{11} W, about 2.7% of the total tidal dissipation but of little consequence to the present thermal state of the Earth. However, it could have exceeded 10^{12} W very early in the history of the Earth, when the Moon was closer and the Earth was rotating faster.

The crust is a light differentiate that has separated from the mantle, releasing gravitational energy in the process. We consider only the continental crust, which has progressively accumulated, and not the oceanic crust, which is continuously recycled back into the mantle. The continental crust, volume $V = 8 \times 10^{18}\,\mathrm{m}^3$, is less dense than the underlying mantle by an average $\Delta\rho = 470\,\mathrm{kg\,m}^{-3}$, so that $V\Delta\rho = 3.8 \times 10^{21}$ kg. The average gravitational potential through which this mass has moved depends on whether the crust is believed to have been derived uniformly from the whole mantle or only from the upper mantle. We calculate each case and adopt a compromise, because the upper mantle appears to have been more affected. It is convenient to the calculation that the density profile of the Earth happens to give almost constant gravity, $g \approx 10\,\mathrm{m\,s^{-2}}$, over the whole depth of the mantle (see Appendix F). Since at radius r, $g = Gm(r)/r^2$, the mass inside r is

$$m(r) = (g/G)r^2, \tag{21.25}$$

where $G = 6.67 \times 10^{-11}\ \mathrm{m^3kg^{-1}s^{-2}}$ is the gravitational constant, so that, in this approximation, (g/G) is constant. It allows a simple calculation of the average gravitational potential difference between radius r_1 and the surface at r_2, weighted according to the distribution of mass in this range. By Eq. (21.23) the mass in the range r to $r + \mathrm{d}r$ is a fraction of the mass M between r_1 and r_2, given by

$$\mathrm{d}m/M = 2r\mathrm{d}r/(r_2^2 - r_1^2), \tag{21.26}$$

and its gravitational potential relative to the surface is $g(r_2 - r)$, so that the gravitational energy released by separation of the mass difference $V\Delta\rho$ uniformly from the mass in the range r_1 to r_2 is

$$\begin{aligned} E &= \frac{V\Delta\rho g}{(r_2^2 - r_1^2)} \int_{r_1}^{r_2} (r_2 - r) r\,\mathrm{d}r \\ &= V\Delta\rho g \left[\frac{r_2}{3} - \frac{2r_1^2}{3(r_1 + r_2)}\right]. \end{aligned} \tag{21.27}$$

For the two cases, separation from the whole mantle or from the upper mantle, the values of this energy are 4.9×10^{28} J and 1.2×10^{28} J.

Accretion of the continental crust is a consequence of convection, which was faster in the past than at present. We assume that the rate of accretion is proportional to the rate of convective heat transport, calculated in Chapter 23, and that the present rate is half of the average over the life of the Earth. On this basis the present rates of energy release for the two cases are 1.7×10^{11} W and 4.3×10^{10} W. We should probably add an energy of accumulation of dense differentiates in the D$''$ layer, but have no quantitative estimate and presume it to be smaller. In Tables 21.1 and 21.4 we take rounded averages of the alternative estimates.

22

Thermodynamics of convection

22.1 Preamble

We have a good measure of the rate of heat loss from the solid Earth to the atmosphere and oceans, 44.2×10^{12} W (Pollack *et al.* 1993). Although the final stage is observed as conduction through the crust and hydrothermal circulation of sea water through young ocean floor, most of the heat comes from the deep interior. Thermal diffusion in a body the size of the Earth is too slow to have cooled the lower mantle noticeably in its lifetime. The deep heat is transported upwards by hot material, driven by the buoyancy of its thermal dilation relative to the cooler material that sinks to replace it. The global scale motion is evident at the surface as tectonic activity, including continental drift and earthquakes, with volcanism as a side effect. Chapter 13 considers the stresses involved in this process and relates them to the mechanical energy derived from convection. The calculation of the energy is presented in this chapter. It is an application of classical thermodynamics.

The mechanical power of convection is the product of heat transport and a thermodynamic efficiency that we can calculate in two ways. First, we recognize that the mantle is a heat engine in the classical thermodynamic sense and that we can apply the Carnot theorem to derive the mechanical power of convection without an explicit consideration of the forces involved. Then, we relate the conclusion more directly to the driving forces of convection, as outlined in Chapter 13, by calculating the efficiency from the buoyancy forces acting on the rising and falling limbs of a convective cell. The two calculations are complementary. The mechanical model shows not only how convective power is generated, but that its efficiency is that of a Carnot cycle operating over an adiabatic temperature ratio between its heat source and sink. We point out here that this is a completely general result. We are calculating the power that is inevitably generated, not merely may be generated, and show that no more power is possible, in principle. This is a fundamental theorem on which a discussion of tectonics must be based, as well as being an essential consideration in the thermal history of the mantle (Section 23.4).

The thermodynamic theorem is first presented in a special, limited form by applying physical restrictions that make the principle easy to see, and then we show that the restrictions can be removed without affecting the conclusions. We consider a homogeneous layer with an adiabatic temperature gradient. As pointed out in Section 19.5, such a layer is neutrally buoyant and a steeper gradient would be required to make it convect, but we can nevertheless postulate that a parcel of material at the bottom of the layer is given an infinitesimal amount of heat which it carries up and releases at the top. By making the heat infinitesimal we can ignore the thermodynamically irreversible processes of heat conduction into and out of the parcel and treat them as isothermal. Then, by allowing the parcel to return to its original position and temperature, we take it round a Carnot cycle between two adiabats and two isotherms. In fact, the return path is not really necessary because the parcel considered and the material it displaces by its return are identical. So, we can appeal to the standard theorem that the mechanical energy, W, generated by the cycle, being the area of a loop in a P–V diagram, is the heat input, Q_1, multiplied by a thermodynamic efficiency, η, given by

$$\eta = \frac{W}{Q_1} = \frac{T_1 - T_2}{T_1} = 1 - \frac{T_2}{T_1}, \tag{22.1}$$

where T_1 and T_2 are the temperatures of the heat source and sink at the bottom and top of the layer (assumed to be the equilibrium temperatures at those levels). By restricting the applied temperature increment to an infinitesimal value, and the temperature profile of the medium to be precisely adiabatic, we can adopt the ideal Carnot efficiency, given by Eq. (22.1). In the following section we show how this is generalized to arbitrary temperature increments and gradients.

22.2 Thermodynamic efficiency, buoyancy forces and convective power

Conservation of energy (the first law of thermodynamics) means that the heat exhausted to the sink is

$$Q_2 = Q_1 - W, \tag{22.2}$$

so that we can write Eq. (22.1) in a less familiar form,

$$\eta' = \frac{W}{Q_2} = \frac{T_1 - T_2}{T_2} = \frac{T_1}{T_2} - 1 = \frac{\eta}{1-\eta}. \tag{22.3}$$

A reason for interest in η', the efficiency referred to the heat exhausted to the sink, rather than the conventional η, referred to the source, is that in the case of the Earth we observe the heat exhausted at the surface. The mechanical power generated is consumed in deforming and so heating the mantle, being thereby put back into the heat source, although not necessarily distributed in the same way, and effectively reduces the net heat input to Q_2. Note that it is not the difference in the temperatures of the heat source and sink that appears in the efficiency, it is their ratio.

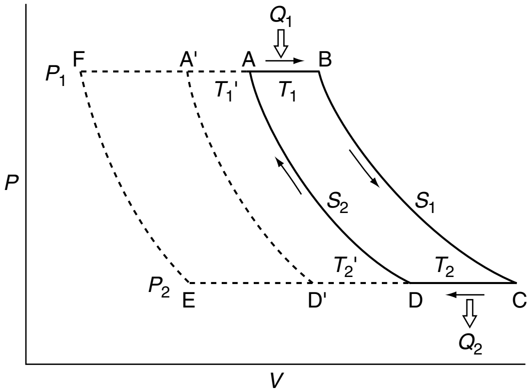


FIGURE 22.1 A thermodynamic cycle between pressures P_1 and P_2. The solid loop is of infinitesimal width, involving infinitesimal heat input, Q_1, on limb AB and output, Q_2, on limb CD. The temperature changes on these limbs are, therefore, infinitesimal and they can be treated as isothermal. The other two limbs are adiabats of entropies S_1 and S_2 (also infinitesimally different), making the loop ABCDA a Carnot cycle. When an adjacent cycle, A′ADD′A is added, the limb DA is cancelled, but temperatures T_1' and T_2' cannot be treated as equal to T_1 and T_2. They are, however, adiabatically related to each other in the same way as T_1 and T_2. This allows a finite extension of the loop to EF, with limbs FB and CE involving finite temperature changes, but with each point on FB adiabatically related to a point on CE by the same pressure difference. The mechanical work done on one cycle is the area of the loop on the P–V diagram.

We can extend the loop with an indefinite number of adjacent loops, each of which is a Carnot cycle, as in Fig. 22.1. The limbs at pressures P_1 and P_2 then involve finite temperature changes and cannot be regarded as isotherms, but each point at P_1 is adiabatically related to a point at P_2 and gives an efficiency corresponding to that section of the total loop. This is given by the temperature ratio, calculated by Eq. (19.19), which integrates to give

$$\ln\left(\frac{T_1}{T_2}\right) = \int_{P_2}^{P_1} \frac{\gamma}{K_S}\,\mathrm{d}P = \int_{\rho_2}^{\rho_1} \frac{\gamma}{\rho}\,\mathrm{d}\rho. \tag{22.4}$$

In general γ and K_S are temperature dependent so that these component efficiencies are not all equal, but the temperature variations in physical properties are quite small, especially at high pressure, and there is no difficulty in calculating the average efficiency for the complete loop. Even a large temperature difference between B and F (or C and E) in Fig. 22.1 does not compromise the efficiency argument. Of course, to get heat into the medium on limb FB the source may need to be at a higher temperature, and to get heat out on limb CE the sink must certainly be at a lower temperature, but these excess temperature differences contribute nothing to the convective cycle per se. They merely drive thermal diffusion, which is thermodynamically irreversible and cannot affect the calculated power or efficiency. Convective power is determined by the adiabatic temperature ratio, whatever higher ratio actually drives the convection.

Now we can also ask whether it is necessary for limbs BC and DA in Fig. 22.1 to be adiabats. If they are not, then it means that heat leaks into or out of the medium on these limbs. If it leaks from one limb of the cycle to the other then it is simply returned to its source and contributes nothing to either heat transfer or convective power. If it leaks into some non-participating part of the medium then we do not have a closed cycle and additional material must be brought into the energy balance. Inevitably some of the heat from the source, or equivalently coolness from the sink, gets only part way round a cycle. But, if we disallow time variations of the thermal structure, this only means that part of the heat participates in a shortened cycle, with its own thermodynamic efficiency. This is directly relevant to mantle convection for which it is convenient to describe convection in terms of coolness transported downwards by subducting slabs, rather than heat transported upwards. The assumption of a uniform distribution of heat sources requires that the subducting coolness is distributed through the mantle and does not all end up at the bottom.

We now examine a mechanical model that is more obviously related to the forces driving convection in the Earth. The cycle ABCDA in Fig. 22.2 is the physical path of a mass m between depths z_1 and z_2 and corresponds to the elementary cycle shown by the solid line in Fig. 22.1. On limb AB of the path it absorbs (or acquires by its own radioactivity) heat Q_1, so that the temperature rises from T_A to T_B and

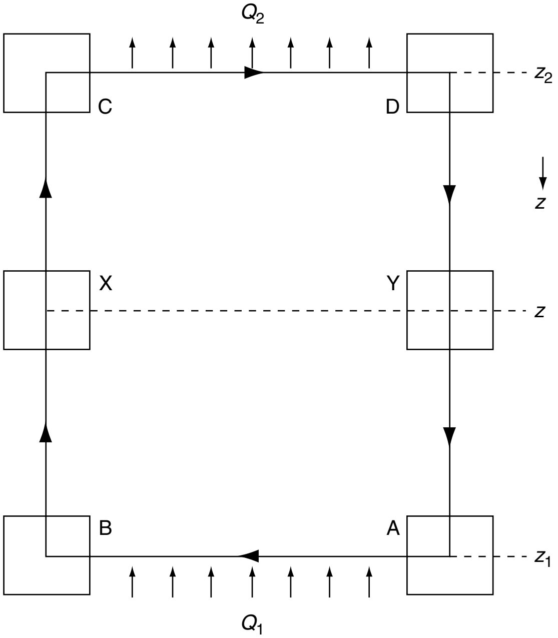


FIGURE 22.2 The physical path of an element of material in a convective cycle.

$$Q_1 = mC_{P1}(T_A - T_B). \tag{22.5}$$

As before, we assume initially that $(T_B - T_A)$ is infinitesimal, but we allow the surrounding medium to have any arbitrary temperature profile. The assumption of an infinitesimal temperature increment means that physical properties, specific heat, C_P, volume expansion coefficient, α, and Grüneisen parameter, γ, are taken as independent of temperature over the small increment, although they may vary arbitrarily with pressure. The mass element then rises from B to C, cooling to temperature T_C, and we make the further assumption, removed later, that this limb of the path is adiabatic. It loses heat on limb CD, cooling to temperature T_D, which is so adjusted that with adiabatic recompression on limb DA it returns to temperature T_A.

The buoyancy forces at arbitrary depth z, at positions X and Y in Fig. 22.2, are

$$F_X = mg\alpha(T_X - T), \tag{22.6}$$

$$F_Y = mg\alpha(T_Y - T), \tag{22.7}$$

where g is gravity and the surrounding ambient temperature, T, is assumed to be the same at X and Y. Then, since no mechanical work is done on the horizontal limbs AB and CD, the net work done by buoyancy forces over the whole cycle is

$$W = \int_{z_2}^{z_1} (F_X - F_Y)\mathrm{d}z = m \int g\alpha(T_X - T_Y)\mathrm{d}z. \tag{22.8}$$

This result does not depend on T and it is the same whether the convection is considered to be driven by upwelling at X or subduction at Y. Since BC and DA are adiabats the entropy difference, ΔS, between any point on one and any point on the other is the same. Thus

$$\begin{aligned}\Delta S &= \int \frac{\mathrm{d}Q}{T} = \int_{T_A}^{T_B} \frac{mC_{P_1}}{T}\mathrm{d}T = \int_{T_Y}^{T_X} \frac{mC_P}{T}\mathrm{d}T \\ &= \int_{T_D}^{T_C} \frac{mC_{P_2}}{T}\mathrm{d}T\end{aligned} \tag{22.9}$$

and, with the assumption of an infinitesimal temperature difference,

$$\begin{aligned}\Delta S &= mC_P \ln(T_X/T_Y) \approx mC_P(T_X/T_Y - 1) \\ &= mC_P \ln(T_A/T_B) \approx mC_P(T_A/T_B - 1) \\ &= Q_1/T_A.\end{aligned} \tag{22.10}$$

Using this result to substitute for $(T_X - T_Y)$ in Eq. (22.8),

$$W = \frac{Q_1}{T_A}\int_{z_2}^{z_1} \frac{g\alpha T_Y}{C_P}\mathrm{d}z = \frac{Q_1}{T_A}\int_{z_2}^{z_1} \frac{\gamma\rho g T_Y}{K_S}\mathrm{d}z. \tag{22.11}$$

The integrand in Eq. (22.11) is the adiabatic temperature gradient (see Eqs. (19.55) and (19.56)), so we can write Eq. (22.11) in a form coinciding with the conclusion of Section 22.1:

$$\begin{aligned}W &= Q_1\frac{T_A - T_D}{T_A} = Q_1\left(1 - \frac{T_D}{T_A}\right) \\ &= Q_1\left\{1 - \exp\left[\int_{\rho_1}^{\rho_2} \gamma\frac{\mathrm{d}\rho}{\rho}\right]\right\} \\ &\approx Q_1\left[1 - \left(\frac{\rho_2}{\rho_1}\right)^{\gamma}\right].\end{aligned} \tag{22.12}$$

The efficiency, $\eta = W/Q_1$, is calculated knowing only the density variation and the Grüneisen parameter. It is not necessary to know the temperatures. η depends only the adiabatic temperature ratio between the heat source and sink. This is given by Eq. (22.4) and does not depend on the actual temperatures unless the thermal properties are temperature dependent.

As mentioned, the assumption of an infinitesimal temperature increment is removed by considering a large number of infinitesimally displaced cycles and allowing for the temperature variations of properties by taking the average efficiency over any temperature range. If we postulate that limb BC or DA departs from adiabats by leakage of heat to or from the surrounding medium, then we must include additional material in a more complicated cycle to maintain a closed system, and if heat leaks from X to Y it is merely returned to the source at A and contributes nothing. The rising and sinking limbs of a convective cycle do not need to be adiabats. The essential conclusion is that the thermodynamic efficiency of convection, that is, the ratio of mechanical power generated to the convective heat transport, is the Carnot efficiency corresponding to the adiabatic temperature ratio of the heat source and sink. Excess temperature gradients are required to make convection occur, but they do not yield more power. They drive the thermodynamically irreversible process of conduction, by which heat enters and leaves the convecting medium. This makes numerical calculations simple because we need only the thermal properties in Appendix G and not the absolute temperatures. However, before discussing numerical results, we consider a complication arising from phase transitions.

22.3 Convection through phase transitions

At depths between 200 km and 700 km, mineral phase transitions complicate convection in ways that have been subject to much debate. The transitions obey the Clausius–Clapeyron equation, which is considered in the special case of melting in Section 19.4 (Eq. (19.40)), but is quite general. We apply it here to the solid–solid transitions of mantle minerals, whose crystal

Table 22.1 Characteristics of mantle phase transitions (olivine composition)

Depth (km)	220	410	660
$\Delta\rho$ (kg m^{-3})	212	94	301
ΔS (J K^{-1}kg^{-1})	−40	−35	+49
ΔT (K) (Eq. (22.14))	+61	+54	−79
dT_C/dP (K MPa^{-1})	+0.44	+0.21	−0.36
$\delta z/\delta T$ (m K^{-1}) (Eq. (22.15))	+69	+135	−70
σ (MPa) ($\delta T = 200$K)			39
Δz (km) (Eq. (22.24))	−4.3	−7.6	+5.4
v(critical)(cm/year) (Eq. (22.29))	1.47	0.83	1.17

structures adjust to higher pressure forms at depth. We write the variation with pressure of the temperature of any transition, T_C, in terms of the volume and entropy changes of the transition by the Clausius–Clapeyron equation:

$$\frac{\mathrm{d}T_C}{\mathrm{d}P} = \frac{\Delta V}{\Delta S}. \qquad (22.13)$$

The principle of LeChatelier demands that, for any phase transition caused by increasing pressure, the volume change, ΔV, is negative but the entropy change, ΔS, may have either sign, and both positive and negative values are observed for mantle minerals. If ΔS is positive, that is the higher pressure form has the higher entropy, then the mineral absorbs heat in converting to this form. The transition is referred to as endothermic and if external heat is not supplied then the mineral is cooled. The phase boundary at 660 km depth in the mantle is due to an endothermic transition. Conversely, with positive ΔS, we have an exothermic transition of the kind occurring at 220 km and 410 km.

If convection proceeds adiabatically through a transition, with no thermodynamic irreversibility or thermal diffusion resulting from the temperature change, then Eqs. (22.1) and (22.12) still apply, where $(T_1 - T_2)$ or $(T_A - T_D)$ now include the temperature increment or decrement of the phase transition. An exothermic transition increases the adiabatic temperature ratio between the heat source and sink, increasing the thermodynamic efficiency of the convection, whereas an endothermic transition reduces the temperature ratio and the efficiency. The 660 km transition has received particular attention because of its inhibiting effect on whole mantle convection. Also, the assumption of perfect thermodynamic reversibility and neglect of thermal diffusion are unrealistic, and all such irreversible effects reduce the efficiency for all transitions, whether endothermic or exothermic. The effect of thermal diffusion in this situation is considered in Section 22.5.

Table 22.1 gives basic numerical data for the transitions that we must consider. The depths and absolute densities as well as values of bulk modulus that are needed for calculations are taken from PREM, but the density increments, $\Delta\rho$, do not coincide with the PREM values, being estimated from mineralogical data, along with the entropy increments, ΔS. Other thermal properties are taken from Appendix G. ΔT is the temperature increment corresponding to ΔS, that is

$$\Delta T = T\Delta S/C_P. \qquad (22.14)$$

It is a notional quantity, not directly observable, but its significance is that it adds to or subtracts from the adiabatic temperature difference $(T_A - T_D)$ in Eq. (22.12). Its magnitude, relative to the total adiabatic temperature range of the mantle, about 1100 K, is a measure of the importance of phase transitions to the total mechanical energy generated by convection (5% to 8%, see Table 22.1). However, the total energy does not determine what happens locally. The detail of what happens is of greatest interest in the case of the endothermic 660 km transition, for which ΔT is negative, because this presents an obstacle to convection through it.

Also listed in Table 22.1 is the Clapeyron slope, dT_C/dP, given by Eq. (22.13). From this we obtain the ‘overshoot’, δz, that is, the difference between the depths at which the transition occurs in upgoing and downgoing material, according to their temperature difference, δT. In the table this is given as a ratio, that is, the depth of overshoot per degree temperature difference,

$$\frac{\delta z}{\delta T} = \frac{1}{\rho g \mathrm{d}T_C/\mathrm{d}P}. \quad (22.15)$$

Bina and Helffrich (1994) discuss the topography of mantle phase boundaries that result from this effect. For subducting material, which is sinking because it is cooler and denser, δT is negative, so the 660 km phase boundary, for which $\mathrm{d}T_C/\mathrm{d}P$ is negative, is depressed by 70 m per degree temperature difference. This means 14 km if the subducting slab is 200 K cooler, on average, than its surroundings. It pushes (or pulls) lower density material into the higher density phase, producing a buoyancy force, $(g\delta z\Delta\rho)$ per unit cross section of slab, opposing its subduction. To estimate the stress we must subtract the driving force of thermal shrinkage, $g(\alpha\rho\delta T)\delta z$ per unit cross section, where α is the volume expansion coefficient. For a temperature contrast δT the net force per unit area of the slab cross section is therefore

$$\sigma = \frac{\delta T}{(\mathrm{d}T_C/\mathrm{d}P)}\left(\frac{\Delta\rho}{\rho} - \alpha\delta T\right), \quad (22.16)$$

where $\Delta\rho$ is the density increment of the transition. Note that the second term in Eq. (22.16) depends on the square of the temperature decrement, δT, because the thermal contraction and the depth of penetration are each proportional to δT. However, the first term is dominant for any plausible value of δT.

We can refer to σ, as calculated by Eq. (22.16), as the virtual stress. It is a rough estimate of the boundary stress caused by subduction through an endothermic transition, but the actual stress depends on geometrical factors, especially the slab thickness relative to the depth overshoot. If the initial maximum temperature deficit of subducting material is ~1000 K, with an average of 500 K, diffusion of heat from the surrounding mantle produces a thicker slab with $\delta T \approx 200$ K. A value of stress, σ, for $\delta T = 200$ K is given in Table 22.1 for the 660 km transition. Even acknowledging that we have probably overestimated the stress by assuming a 100% olivine composition of the subducting material, this stress exceeds the average tectonic stress estimated in Section 13.2. Thus, subduction through this transition presents a mechanical problem. Seismological evidence of horizontal deflections of slabs and accumulations of subducted material at the boundary is referred to in Chapter 12, but some slabs appear to pass straight through. Numerical simulations (e.g. Solheim and Peltier, 1994; Tackley *et al.*, 1994) have suggested that accumulated material may avalanche into the lower mantle after a delay.

22.4 Thermodynamic efficiency of mantle convection and tectonic power

The mechanical power, $\dot{E}$, generated by the convection of heat, $\dot{Q}$, from a source at pressure P_1 to a sink at pressure P_2 is given by rewriting Eq. (22.12) in a more general way:

$$\dot{E} = \dot{Q}\int_{P_2}^{P_1} (\partial \ln T/\partial P)_S \,\mathrm{d}P = \dot{Q}\int_{P_2}^{P_1} (\gamma/K_S)_S \,\mathrm{d}P. \quad (22.17)$$

We use this to calculate the tectonic power from the distribution of heat sources. It is not necessary to know the temperatures, but only adiabatic ratios that are expressed in terms of physical properties. The intervention of phase transitions is handled by introducing discontinuities in the adiabatic temperature profile according to the entropies of the transitions in Table 22.1:

$$\Delta T/T = \Delta \ln T = -\Delta S/C_P. \quad (22.18)$$

There are several ways of representing the results of such calculations. A straightforward calculation of the thermodynamic efficiency of convective heat transport between any two depths is simply $\eta = \dot{E}/\dot{Q}$ by Eq. (22.17) and this is the fundamental result. Figure 22.3 is a plot of this efficiency for convection of heat to the surface as a function of the depth from which it originates. The efficiency corresponding to the transport of heat between any two depths is the difference between the values at those depths. The calculations assume that convective heat transport stops at the base of the lithosphere and that heat loss from the lithosphere to the

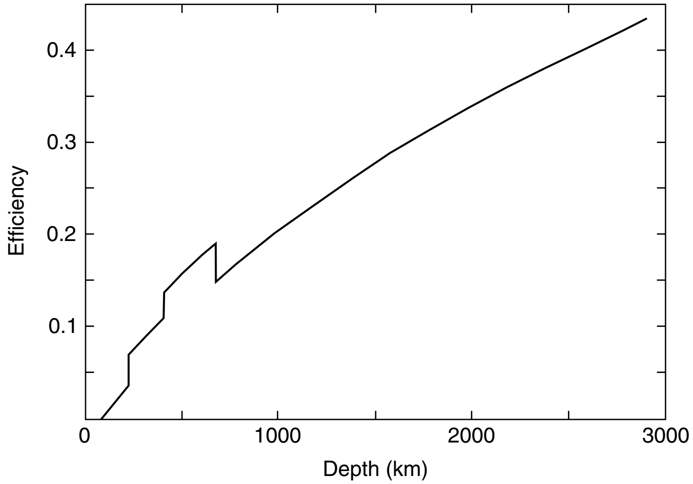


FIGURE 22.3 Thermodynamic efficiency of convective heat transport to the surface from depth *z* in the mantle as a function of *z*, neglecting irreversibility of phase transitions.

surface is by conduction and makes no contribution to tectonic power.

The power calculation requires an assumption about the distribution of heat sources. For this purpose radiogenic heat and net cooling are treated slightly differently. We assume radiogenic heat to be distributed uniformly, so that the heat generation per unit volume, $\dot{q}_R$, is proportional to density,

$$\dot{q}_R = \dot{Q}_R \rho / M_M = 5 \times 10^{-12}\,\mathrm{W\,kg^{-1}} \times \rho, \tag{22.19}$$

where, by the preferred model in Chapter 21, $\dot{Q}_R = 20 \times 10^{12}$ W is the total radiogenic heat in the mantle, mass $M_M = 4 \times 10^{24}$ kg. The heat loss by cooling is proportional to temperature because the temperature profile is assumed to be adiabatic, with temperatures having constant ratios as they decrease. This means that, as in the calculation of the effective heat capacity (Section 21.3), the rate of heat loss from unit volume at temperature T is calculated from the rate of change of the potential temperature, T_P,

$$\dot{q}_C = 1.195 \rho C_P \frac{T}{T_p} \frac{\mathrm{d}T_P}{\mathrm{d}t}. \tag{22.20}$$

The factor 1.195 is the ratio of Eqs. (21.7) and (21.5), arising from the gravitational contribution to the effective heat capacity. For the temperature ratio we use the adiabatic one, as in Eq. (22.17),

$$\frac{T}{T_P} = \exp\left[\int_0^P (\gamma/K_S)\mathrm{d}P\right]. \tag{22.21}$$

The cooling rate is

$$-\frac{\mathrm{d}T_P}{\mathrm{d}t} = \frac{\dot{Q} - \dot{Q}_R}{\phi_m} = 1.69 \times 10^{-15}\,\mathrm{K\,s^{-1}}$$
$$= 53\ \mathrm{K}/10^9\ \mathrm{years}, \tag{22.22}$$

where $\dot{Q} = 32.5 \times 10^{12}$ W is the total convected heat loss, $\dot{Q}_R = 20 \times 10^{12}$ W is the radiogenic heat of our model, as in Section 23.4, and $\phi_m = 7.4 \times 10^{27}\,\mathrm{J\,K^{-1}}$ (Eq. (21.7)). The heat flux through any depth is the integral of $(\dot{q}_R + \dot{q}_C)$ for the volume below that depth.

Integration of Eq. (22.17) gives the total power of mantle convection. Assuming that all the heat is convectively transported, this is $E = 8.03 \times 10^{12}$ W. There is a correction for the fact that the conducted heat is not zero, although it is small. By calculating the power that would be generated by the conducted heat at each level if it were convected, and integrating over the whole mantle, we obtain the mechanical power loss due to conduction, 0.3×10^{12} W. Subtracting this from the power estimated with neglect of conduction, we have a net convective (tectonic) power of 7.7×10^{12} W. This power is not generated uniformly through the mantle. Its distribution is calculated by taking the incremental thermodynamic efficiency for a radius range Δr, noting that, by differentiating Eq. (22.1), $\mathrm{d}\eta = \mathrm{d}\ln T$,

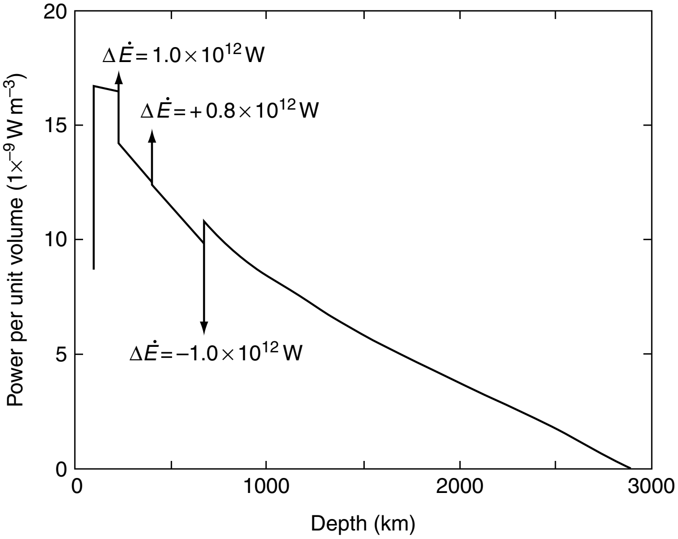


FIGURE 22.4 Convective power generation per unit volume of the mantle as a function of depth, showing its concentration in the upper mantle. The three phase transitions appear as discontinuities that can be regarded as delta functions in this figure, two positive and one negative. Their contributions to the total convective power are indicated.

$$(\mathrm{d}\eta/\mathrm{d}r)\Delta r = (\partial \ln T/\partial P)_S(\mathrm{d}P/\mathrm{d}r)\Delta r$$
$$= (\gamma\rho g/K_S)\Delta r, \qquad (22.23)$$

and multiplying by the heat flux through that radius. Figure 22.4 shows the result of this calculation, represented as power generation per unit volume. The essential feature is the general decrease in power generation with depth. This is a consequence of the fact that power generation at any depth is proportional to the heat flux and all of the deep heat is convected through the shallow layers. Although there can be no precise correspondence between the generation and dissipation of convective power, there must be a general correspondence and this necessarily means that convection is less vigorous in the lower mantle than in the upper mantle. However, the whole lower mantle must be involved because to lose heat it must be convecting, but its convection is slower and this is consistent with a higher viscosity.

The effect of the phase transitions on thermodynamic efficiency is shown in Fig. 22.3 and their contributions to convective power are indicated in Fig. 22.4. Although the negative transition at 660 km depth has a greater entropy than the shallower transitions, the heat flux through it is less, reducing its effect on total power, relative to the two positive transitions.

Referred to the heat flux to the surface, the total convective power represents an efficiency of 24%. This is higher than might be expected from a glance at Fig. 22.3. The reason is that the heat per unit volume from the lower mantle is enhanced by the factor $\rho T/T_\mathrm{P}$ in Eq. (22.20). But, in spite of the dominance of lower mantle heat, the power generation in the lower mantle is much weaker than in the upper mantle, because the heat is transported over a small temperature ratio within the lower mantle itself. Conversely, all of the lower mantle heat passes up through the entire upper mantle. If the upper mantle were convecting alone, with no heat from the lower mantle, the power generation would be 0.83×10^{12} W, little more than 10% of the power of whole mantle convection and insufficient to account for the tectonic dissipation discussed in Section 13.2. With lower mantle heat, the upper mantle convective power is 5.3×10^{12} W, obtained by integrating the convective power per unit volume, plotted in Fig. 22.4.

22.5 Why are mantle phase boundaries sharp?

Two observations on phase boundaries offer clues to the manner of mantle convection: (i) over most of the mantle, that is except in the limited areas of strong subduction, the depths of the boundaries are more or less uniform (say ±10 km), and (ii) the boundaries are sharp enough to reflect short period seismic waves, indicating that the depth ranges over which the transitions at 410 km and 660 km occur are less than 2 km (Xu *et al.*, 2003). The implication of observation (i) is that subduction occurs in narrow bands and that the return flow is a broad upwelling over the rest of the mantle. Temperature contrasts between upgoing and downgoing material are presumed to be several hundred degrees and, by the values of $\delta z/\delta T$ in Table 22.1, this means boundary deflections of tens of kilometres. Thus there is no evidence of differential motion over most of the mantle and it must all be doing the same thing – rising. The subduction of cool slabs is compensated by a broad-scale return flow.

Observation (ii), that the boundaries are sharp, confirms that over most of the mantle the vertical motion is slow, less than 1 cm per year, compared with the plate speeds of several centimetres per year. There are two reasons why thicker phase boundaries might be expected. As Solomatov and Stevenson (1994) pointed out, under isothermal (equilibrium) conditions a single mineral may undergo a phase transition at sharply defined P–T conditions but this cannot be true for the multi-component mineral mix of the mantle. For the 660 km transition the expected depth spread is nevertheless quite small, and Solomatov and Stevenson addressed more particularly the 410 km transition, for which equilibrium thermodynamics predicts a depth range much greater than 2 km. They show that the narrowing of the transition can be explained as a consequence of the fact that phase transitions often do not occur under precisely equilibrium conditions but require nucleation under conditions of metastable overshoot. For the 410 km transition they estimate this to correspond to about 10 km in depth, which would sharpen the transition sufficiently to give the observed seismic reflections. Such an overshoot would occur for transitions in both directions and reduce by about 1% the thermodynamic efficiency calculated in Section 22.2.

The other mechanism that causes thickening of phase boundaries is directly relevant to our ideas about the pattern of mantle convection. It applies even to the transformation of a simple single component if it occurs adiabatically, that is too rapidly to allow isothermalization by thermal diffusion. This is illustrated in Fig. 22.5 for the case of an endothermic transition, as at 660 km. Consider material subducting on an adiabat and meeting the boundary at A. As soon as it does so it begins to transform to the higher density phase, but this causes cooling and the transformation goes to completion only when a further increase in pressure suffices to overcome the temperature drop, ΔT, at point B. As long as both phases, are present the material follows the Clapeyron path AB, and at greater depths it resumes the normal adiabatic gradient. Equation (22.14) gives ΔT, as listed in Table 22.1, and the consequent depth range, Δz, is given by

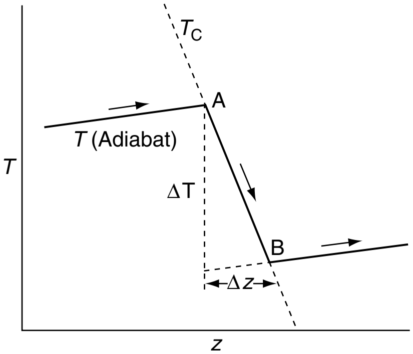


FIGURE 22.5 Temperature–depth relationship for rapid (adiabatic) subduction through the transition at 660 km depth, for which the transition temperature, T_C, decreases with pressure or depth.

$$\Delta z = \Delta T/\left[-\frac{\mathrm{d}T_C}{\mathrm{d}z} + \left(\frac{\partial T}{\partial z}\right)_S\right]. \qquad (22.24)$$

Note that $\mathrm{d}T_C/\mathrm{d}z$ is negative for the case illustrated, but $(\partial T/\partial z)_S$ is always positive (unless one considers materials with negative expansion coefficients). The depth range over which this occurs, as illustrated in Fig. 22.5, gives Δz, as listed in Table 22.1. Since this is greater than the transition zone thicknesses indicated by seismology, we appeal to thermal diffusion to equalize temperature and sharpen the transitions.

When thermal diffusion is allowed the temperature profile takes the form illustrated in Fig. 22.6, in which the phase transition is restricted to the range CD. Outside this range

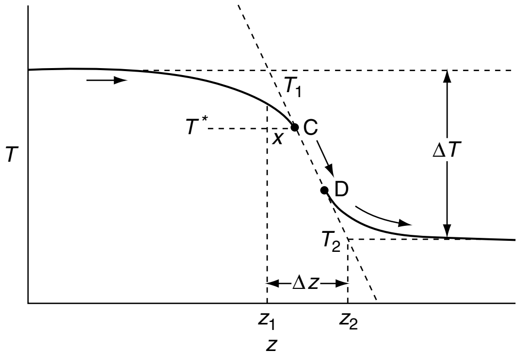


FIGURE 22.6 Effect of thermal diffusion at the 660 km phase boundary. The small ambient adiabatic temperature gradient is ignored here, so that the remote temperature limits are T_1 and T_2.

we add a term to the thermal diffusion equation to account for the motion, at speed v. Ignoring the small adiabatic temperature gradient,

$$\frac{\partial T}{\partial t} = \eta \frac{\partial^2 T}{\partial z^2} - v \frac{\partial T}{\partial z}, \tag{22.25}$$

where $\eta = \kappa/\rho C_P$ is thermal diffusivity. Equation (22.25) represents the temperature at a fixed point, with the material moving past, carrying its temperature with it and so counteracting the diffusion and producing a steady state in which $\partial T/\partial t = 0$ at any level. Then, for the approach to point C we can integrate Eq. (22.25) to give

$$\eta \frac{\partial T}{\partial z} = v(T - T_1), \tag{22.26}$$

where boundary conditions $\partial T/\partial z = 0$ and $T = T_1$ at $z \to -\infty$ are applied. As long as C and D do not meet, the boundary conditions at C, depth $(z_1 + x)$, are $\partial T/\partial z = \mathrm{d}T_C/\mathrm{d}z$ and $T = T^*$, where

$$T^* = T_1 + \left(\frac{\mathrm{d}T_C}{\mathrm{d}z}\right)x. \tag{22.27}$$

Introducing this boundary condition to Eq. (22.26), we have

$$\eta = vx. \tag{22.28}$$

The same applies to the lower bound of the transition at D, which is displaced upwards by distance x from z_2. Thus, if $x = (z_2 - z_1)/2 = \Delta z/2$, the layer shrinks to zero thickness. This applies to speeds

$$v \leq 2\eta/\Delta z. \tag{22.29}$$

With the values of Δz in Table 22.1 and assuming diffusivity $\eta = 1.0 \times 10^{-6}\,\mathrm{m}^2\,\mathrm{s}^{-1}$, Eq. (22.29) gives the listed values of v(critical), the convective speeds below which thermal diffusion makes the boundaries sharp. It is evident that the sharpness of phase boundaries, outside the more confused subduction zones, imposes a limit on convective speeds significantly lower than the speeds of most of the surface plates. This conclusion coincides with the convective pattern favoured by the energy argument in Section 13.2.

22.6 Compositional convection in the core

Compositional convection was first recognized to be important to core energetics and the dynamo by S.I. Braginsky and confirmed by Gubbins (1977) and Loper (1978a,b). In Section 21.4 it is referred to as one of the components of the core heat budget, but that is almost incidental to its role in the dynamics and evolution of the core. We give it a closer examination here. In particular, we seek a simple analytical treatment of this important effect. We make the conventional assumption that the light core solute rejected by the solidifying inner core mixes uniformly into the outer core, releasing gravitational energy in the process The possibility that this assumption is not completely valid is considered briefly in Section 23.5. In Section 21.4 the energy is determined from models of the whole Earth with and without an inner core. The value so calculated, 5.63×10^{28} J, includes elastic strain energy and a small component released in the mantle that must be subtracted, leaving 4.79×10^{28} J as the energy relevant to the dynamo. It recognizes that the mass rearrangement modifies the gravity profile by increasing the central density and causing a contraction of the whole Earth, although the energy released in the mantle by this process is less than 5%.

Nevertheless, we are interested to know how big a fraction of the core energy release is attributable to relative motion of materials with different densities and not to the general contraction. With the assumption that the material moves in a fixed gravity profile a simple but realistic analytical calculation is possible.

As in Section 21.3, we take the compositional component of the density contrast between inner and outer cores to be 620 kg m^{-3}. In the volume of the inner core this represents a mass excess of 4.49×10^{21} kg, referred to the outer core density. The fraction of this that has been drawn from the outer core is 0.949, being the outer core fraction of the total mass of the core, so that the mass drawn from the outer core and deposited in the inner core is

$$\Delta m = 4.26 \times 10^{21} \text{ kg}. \tag{22.30}$$

This is gravitationally equivalent to a mass deficiency going the other way. We assume that the inner and outer cores remain compositionally homogeneous, so that this mass is drawn from the outer core in proportion to the local density. Similarly, its deposition in the inner core is proportional to local density.

To calculate the energy released by this process we need the gravity profile. This is reasonably represented by a power law,

$$g = g_R (r/R)^x. \tag{22.31}$$

A least-square fit to the PREM outer core tabulation gives $x = 0.8436$, with $g_R = 10.78$ m s^{-2} at the core–mantle boundary radius, R. We use this relationship in what follows, recognizing that it is a simple analytical approximation and is not exact but greatly simplifies the calculation. It allows us to write the mass inside radius r as

$$\begin{aligned} m(r) &= gr^2/G = (g_R/G)r^{x+2}/R^x \\ &= M_C (r/R)^{x+2}, \end{aligned} \tag{22.32}$$

where M_C is the total core mass. We need the fraction of this mass in the range r to $(r + \mathrm{d}r)$,

$$\mathrm{d}m = (x+2) M_C r^{x+1}/R^{x+2} \mathrm{d}r\,, \tag{22.33}$$

because this is the fraction of Δm (Eq. (22.30)) that originates in a range dr (between R and the inner core radius, r_i). We calculate first the energy released by depositing Δm on the inner core boundary. For this we need the gravitational potential difference between r and r_i,

$$V = \int_{r_i}^{r} g\ \mathrm{d}r = \frac{g_R}{R^x}\int_{r_i}^{\mathrm{r}} r^x \mathrm{d}x = \frac{g_R}{R^x}\left(r^{x+1} - r_i^{x+1}\right)/(x+1). \tag{22.34}$$

Then the energy release is

$$\begin{aligned} E_1 &= \Delta m \int V \frac{\mathrm{d}m(r)}{M_C} \\ &= \Delta m \frac{g_R}{R^x}\frac{(x+2)}{(x+1)}\int_{r_i}^{R}\left(r^{x+1} - r_i^{x+1}\right)\frac{r^{x+1}}{R^{x+2}}\mathrm{d}r \\ &= \Delta m g_R \frac{(x+2)}{(x+1)}\left[\frac{R}{2x+3} - \frac{r_i^{x+1}}{R^x(x+2)} \right. \\ &\qquad\qquad \left. + \frac{r_i^{2x+3}}{R^{2x+2}}\frac{x+1}{(x+2)(2x+3)}\right] \\ &= 9.45 \times 10^6 \Delta m = 4.03 \times 10^{28}\ \mathrm{J}. \end{aligned} \tag{22.35}$$

The second stage of the integration is to distribute Δm through the inner core volume, but it is useful to see Eq. (22.35) separately because for an incremental Δm it gives the energy release at the present stage of inner core formation. For the total energy of inner core formation we add a second integral, set up similarly, with

$$\frac{\mathrm{d}m}{M_C} = (x+2)\frac{r^{x+1}}{r_i^{x+2}}\mathrm{d}r, \tag{22.36}$$

giving

$$E_2 = \frac{\Delta m g_R}{2x+3}\frac{r_i^{x+1}}{R^x} = 4.9 \times 10^{27}\ \mathrm{J}. \tag{22.37}$$

By this calculation the total compositional component of the gravitational energy of inner core formation is

$$E = E_1 + E_2 = 4.52 \times 10^{28}\ \mathrm{J}, \tag{22.38}$$

which is seen to be about 6% less than that calculated by the complete theory, after deduction of the strain energy and the mantle energy release. The difference can be attributed to a general contraction associated with the chemical separation, because that is not allowed for in the approximate calculation above.

The compositional separation energy shares with precessional energy a physical distinction from the other core energy sources listed in Section 21.4, being mechanical energy, whereas the others are thermal energies, from which mechanical energy can be derived only with the thermodynamic efficiency of a heat engine. The simple way of identifying mechanical vs thermal sources is to ask: can the processes be reversed, in principle, simply by applying heat to the core? If the answer is yes, then we must classify them as thermal energy. For example, although the gravitational energy released by thermal contraction may appear to be mechanical, if we could reheat the core then the contraction would be reversed, with no mechanical input and no more heat than was extracted by the cooling. Thus we must count the energy released by contraction as thermal. Otherwise we would have a physical system violating the second law of thermodynamics. We can re-examine the compositional energy in this light, having noted that we can identify two contributions to it. Is the 6% that we attributed to general contraction properly counted with the other 94% as mechanical? By the reheating test the answer must be yes. Being part of the separation process it cannot be recovered by applying heat and it can contribute to dynamo power with high efficiency.

22.7 Thermodynamic efficiency of core convection and dynamo power

There is no simple number to represent the efficiency of core convection. We consider the several components of core energy that have different efficiencies and contribute in different proportions according to the rate of core cooling. The first distinction to make is between thermal and compositional convection. We treat the compositional effect as 100% efficient, although some energy may be lost in turbulent mixing. The efficiency of thermal convection depends on the depth distribution of the heat sources, being greatest for the latent heat, which is derived entirely from the inner core boundary (ICB). The heat from cooling of the inner core is also treated as coming from the ICB. We calculate here the thermodynamic efficiencies of the several energy sources discussed in Section 21.4 and apply them to the calculation of convective power, as summarized in Table 21.5. We recognize the distinction between η and η' (Eqs. (22.1) and (22.2)) by calculating η for each process, but in Eq. (21.22) effectively convert the total to η' by allowing for the re-use of ohmic heat. Conduction is treated as a negative component of convective power in the sense of reducing the power that would be generated if all of the heat flux were convected.

The convective energy of each thermal source is calculated by taking the heat originating in the range dr, at temperature T, multiplying by the efficiency of its convective transport to the core–mantle boundary, $(1 - T_{\mathrm{CMB}}/T)$, and integrating over the volume of the outer core. As in the calculation of heat capacity, T_{CMB} is the reference temperature and an adiabatic temperature gradient is maintained so that cooling is everywhere proportional to absolute temperature and the thermal contribution by an element at temperature T is multiplied by T/T_{CMB} relative to its normal heat capacity, ρC_P. The cooling and freezing components of the heat budget are each accompanied by global contraction and consequent release of gravitational energy. Some of this energy is released in the mantle but only the core component is included in Table 21.5. As pointed out in Section 22.6, this energy must be treated as thermal for the purpose of calculating the efficiency with which it generates convective power, and not as mechanical energy.

Radiogenic heat is assumed to be proportional to density, to accord with compositional homogeneity, making it more concentrated at greater depth. This gives it a slightly higher thermodynamic efficiency than a volumetrically uniform source, which is assumed in the case of ohmic heat. Precessional dissipation, being assumed to be 100% efficient, requires no information about how it is converted to ohmic heat. As discussed in Section 22.6, the energy released by compositional convection is calculated by assuming complete compositional mixing, that, is the light solute is distributed according to local density. Since it is released at the inner core boundary, this means a somewhat smaller energy release than if volumetrically uniform mixing were assumed. In Section 23.5 we note the suggestion that the mixing may not be uniform, but that the light ingredient rises to form a gravitationally stable layer at the top of the core, releasing much more energy. Seismological evidence for such a layer would be needed to justify that assumption here.

Although, in a formal thermodynamic sense, conduction is a process of zero efficiency, it is convenient to the present discussion to treat it as having negative efficiency. The efficiency that it would have if the conducted heat were convectively transported gives an implied power which is subtracted from the total power obtained by assuming all heat to be convected. This is a simple way of accounting for the radial distribution of the loss of convective power, as may be inferred from Fig. 21.3, even without allowing for the efficiency factor. The 'inefficiency' of conduction may be considered modified by refrigerator action (Section 22.8), but we treat the refrigerator efficiency (f in Table 21.5) separately. It depends on the depth range over which conducted heat exceeds the total heat flux. For the two models represented in Figs. 21.3 and 21.4, $f = 0.091$ (for $A = 0$) and 0.043 (for $A = 0.2$).

We identify as dynamo energy all of the convective energy that is not either lost by conduction or consumed in refrigerator action. This means that we assume viscosity, η, to be negligible, although there are no useful direct observations. Estimates of η for the core from geophysical observations cover an extremely wide range (Secco, 1995) but are only upper bounds. Estimates based on liquid metal physics (Poirier, 1988; Dobson, 2002) agree that a value below 1 Pa s can be confidently assumed. Assuming this value, we can estimate the viscous dissipation by internal motion at 4×10^{-4} m s^{-1}, as indicated by the geomagnetic secular variation (Section 24.3), in cells of radius 80 km, the smallest viable size suggested in Section 24.2. This corresponds to a shear rate $\dot{\varepsilon} \approx 5 \times 10^{-9}$ s^{-1}. The corresponding dissipation is $\eta\dot{\varepsilon}^2 < 2.5 \times 10^{-17}$ W m^{-3}, compared with $\sim 2 \times 10^{-9}$ W m^{-3} attributed to ohmic dissipation. As we point out in Section 24.7, the kinetic energy of core motion is very small compared with the magnetic energy. The convective forces work directly against the magnetic field and do not establish significant kinetic energy, so, with negligible viscosity, the convective power is converted directly to magnetic field energy with no losses. Thus, we identify the convective energies in Eqs. (21.23) and (21.24) as dynamo energy and divide by the inner core lifetime, τ, to obtain the mean dynamo power for this period, as plotted in Fig. 21.4.

The mean power plotted in Fig. 21.4, and the equations in Section 21.4 on which it is based, ignore the variations with time of the convective driving forces. We now take a closer look at this. The obvious variation is that of the radiogenic heat, which, being attributed to ^{40}K (half life 1.25 billion years), was more than ten times as strong early in the life of the Earth. The dynamo energy of the non-radioactive model ($A = 0$) in Fig. 21.4 is only marginally adequate, and we recognize the dynamo as a robust feature of the Earth, for which marginal viability is unconvincing. We therefore refer to the 0.2 terawatt model as a preferred model. In this case the early radiogenic heat would have been 2 terawatts, which has no dramatic effect on the thermal history calculation in Section 23.5. However, 2 terawatt models, as have been suggested on account of higher assumed conductivity, would release 20 terawatts in the early core and this is difficult to rationalize with the thermal history.

Dissipation by precessional torques would also have been stronger in the past. Although, by our estimate it was never a major contributor to core energy, being mechanical energy it could be almost 100% efficient for dynamo action, although there must be some ohmic dissipation in the mantle. By a calculation presented in Section 24.7, we estimate that it would have provided about 7×10^{10} W to the early dynamo, a partial offset to the variation in compositional convective energy that we now examine.

In the Q column of Table 21.5 compositional separation appears as a significant but far from dominant source of heat, but its high efficiency makes it the biggest contributor to the E column of the table. But the relative proportions of the entries in this column change with time. The rates of energy release by latent heat and compositional separation are proportional to the rate of growth of the inner core volume, but the heat release by general cooling is proportional to the rate of change in temperature. The ratio of these two rates depends on the inner core size. Lister and Buffett (1995) pointed out that the compositional and freezing contributions were smaller when the inner core was small. We show here how this arises and consider the consequences in Section 23.5.

Both the adiabatic and melting point gradients are proportional to the pressure gradient and therefore to gravity. In the inner core range this is almost proportional to radius, as it would be exactly in a uniform sphere. Thus $-\mathrm{d}T/\mathrm{d}r \propto r$, and as the boundary temperature changes with time, t, so the temperature of the boundary varies with its radius as

$$\mathrm{d}T_{\mathrm{ICB}}/\mathrm{d}t \propto - r\mathrm{d}r/\mathrm{d}t. \tag{22.39}$$

But the inner core volume varies as

$$\mathrm{d}V/\mathrm{d}t = 4\pi r^2 \mathrm{d}r/\mathrm{d}t, \tag{22.40}$$

so that

$$\mathrm{d}V/\mathrm{d}t \propto - r\mathrm{d}T_{\mathrm{ICB}}/\mathrm{d}t. \tag{22.41}$$

To a sufficient approximation for this purpose $T_{\mathrm{ICB}} \propto T_{\mathrm{CMB}}$ and we can substitute $\mathrm{d}T_{\mathrm{ICB}}/\mathrm{d}t = (T_{\mathrm{ICB}}/T_{\mathrm{CMB}})\mathrm{d}T_{\mathrm{CMB}}/\mathrm{d}t$ in Eq. (22.41) to identify it with the cooling rate, as considered in the discussion of heat capacity (Section 21.3). The rate at which growth of the inner core contributed to dynamo power is proportional to its radius and would have begun gradually. There was no sudden onset and therefore no discontinuity for which evidence could be sought in the paleomagnetic record.

We need to consider the variation in dynamo power over time in the light of Eq. (22.41). The rate at which the core loses heat to the mantle is controlled by the physical properties and convective processes in the mantle, particularly the D″ boundary layer. Stacey and Loper (1983) argued that it is stabilized by two competing effects: as the mantle cools and stiffens, so it slows all convective processes, but it cools faster than the core and it is the temperature difference that drives the core-to-mantle heat flux. If we accept that it is the heat flux from the core that is controlled in this way and not the rate of change in temperature, then without any change in the heat flux the cooling rate slows as the inner core grows. This occurs irrespective of radioactivity because the effective heat capacity of the core increases with the inner core size. But if there were ever no inner core, then, in the absence of the very efficient convective processes represented by items 3, 4 and 5 in Table 21.5, the dynamo could not have been maintained without a much greater contribution by items 1, 2, 6 and 7 than would be required to maintain a constant heat flux. This problem is pursued further in Section 23.5 with the conclusion that the constant heat flux argument appears satisfactory for the last 2 billion years but not for the earlier period.

In Section 23.5 we refer also to the suggestion that the light solute rejected by the inner core may become concentrated at the top of the core, releasing much more gravitational energy than with uniform mixing. However, a seismological indication that this is so would be needed to make it the basis of a core energy calculation.

22.8 Refrigerator action in the core

In Sections 21.4 and 22.7 (see also Table 21.5) we refer to refrigerator action as a feature of the core energy balance. We now add a note on how this arises and what it means from a thermodynamic perspective. As a simple example consider a large fluid body of negligible viscosity, and with no heat sources, in a container which is perfectly insulating. We introduce a stirring mechanism strong enough to maintain an adiabatic temperature gradient. Upward conduction of heat would tend to reduce this gradient, but since it is maintained by stirring there is a continuous flow of heat. It cannot escape from the top of the container and is carried down, against the temperature gradient, by the stirring action. This is refrigerator action, which requires mechanical power equal to the power that would be generated by upward convection of the same heat. This power would be released as heat in the fluid, causing a general temperature rise.

As Fig. 21.3 indicates, it is possible for conducted heat to exceed the total heat flux in the outermost part of the core and in this case the difference must be carried down mechanically.

This can be accomplished by compositional convection. The energy available for dynamo action is reduced by an amount that depends on how far the heat must be transported, being determined by the temperature range over which it is transported. The ratio of the mechanical power required to the heat transported is the factor f in Eq. (21.22) and Table 21.5. Thus, for the $A = 0.2$ terawatt model in Fig. 21.3, $f = 0.043$, but for the $A = 0$ model, with no radioactivity, the heat must be carried down to greater depths and $f = 0.091$.

23

Thermal history

23.1 Preamble

A thermal history of the mantle can be calculated almost independently of the core. The logic for this is that core heat is carried up through the mantle by narrow, buoyant plumes that have only a weak interaction with the plate tectonic convection process that cools the mantle. The converse is not true. The core is cooled by loss of heat into a thermal boundary layer at the base of the mantle and so depends on the temperature difference between the core and the deep mantle, 100 to 200 km above the boundary, as well as on mantle rheology. The boundary layer must have developed, that is, the mantle must have cooled substantially, before significant core cooling could occur.

Mantle rheology also controls the cooling of the mantle itself, but it is a mutual control. Tozer (1972) drew attention to the fact that the strong dependence of viscosity on temperature (Eq. (10.27)) has a stabilizing effect on both. If the mantle were to become too cool and viscous to convect at the 'normal' rate, convection would slow until radioactive heating caught up. But this does not mean that the heat loss is in equilibrium with the heat source, because the source is not constant. With diminishing radiogenic heat, convection slows down and this means that the mantle is cooling, as is most convincingly demonstrated by a consideration of the heat balance equation (Eq. (23.14)). Application of this equation shows that the rate at which radiogenic heat decreases is an important control on thermal history.

The other basic principle that we apply is referred to in Section 13.3: convection is driven by sources of buoyancy, positive or negative, generated at boundaries. This means that we need a quantitative description of the process of surface heat loss. As discussed in Chapter 20, this is quite different in continental and oceanic areas. Both must be accounted for. The continents move about and conduct heat to the surface, but are otherwise non-participants in convection and act as blankets on the 40% of the surface area that they cover. This is an area of 2×10^{14} m^2, in which we include the submerged continental margins. Of greater interest to the subject of this chapter is the 3.1×10^{14} m^2 of oceanic crust/lithosphere, which is a thermal boundary layer and acts as the exhaust of the mantle heat engine. It loses heat to the sea until it reaches a subduction zone and returns to the mantle, to be replaced by fresh, hot lithosphere at an ocean ridge. Heat is lost from the oceanic lithosphere by a combination of thermal diffusion and circulation of sea water in cracks. These processes vary differently with time. The hydrothermal circulation decreases in importance with increasing age of the lithosphere. This means that, in extrapolating backwards in time to periods of more rapid convection and shorter residence time of oceanic lithosphere, the relatively greater importance of hydrothermal circulation must be recognized. But the two effects are not independent and there is no theory of the total process, so we introduce a simple empirical expression to represent their combined effect as a function of lithospheric age. This is a subject of Section 23.2, in which we show that a simple dimensional analysis indicates that there is no systematic variation in plate size with time or average convective speed (Eq. (23.5)). This is not an intuitively obvious result, but, as we point out in that section, there appears to be no contradictory evidence.

Returning to the problem of rheological control, the basic equation is Eq. (10.27), in which the exponent n is unknown. Favoured values are $n = 1$ (linear rheology) and $n = 3$ (non-linear) and opinions differ over which is the more appropriate value for the mantle. We examine both alternatives, but the problem of choosing a value of n is compounded with the uncertainty over the value of the parameter g, which is a measure of the activation energy of creep. Together they control the cooling rate, that is, the difference between the heat loss and radiogenic heat, so we treat them both as unknown and use the present radiogenic heat, $\dot{Q}_{R0}$, as a controlled variable to investigate the relationship between them and the cooling history.

The continental crust, which is rich in radioactive elements, has grown with time and its separation from the mantle has progressively reduced the mantle content of these elements. We assume that this differentiation process has occurred at a rate proportional to the speed of convection, and therefore to the rate of mantle heat loss, being much slower now than several billion years ago. In integrating the heat balance equation (Eq. (23.14)) backwards in time, we return the continental crust progressively to the mantle. This has two effects. It gives an increase in the mantle radiogenic heat, additional to the time-dependence arising from radioactive decay, and it increases the surface area of oceanic lithosphere by progressively removing the continents.

Now we mention two effects that we do not consider important and do not include in our analysis for reasons that we explain: the loss of mantle volatiles and latent heat of lava solidification at ocean ridges. The significance of volatile loss is that volatiles reduce viscosity (Fig. 2.3). If this were important to mantle rheology it could be allowed for by varying the activation energy for creep in Eq. (10.22) by making the parameter g a function of time or of integrated heat loss. Such a calculation reverses the thermal history: the mantle would be heating up. If this effect exists it must be exceedingly small. From the water contents of hot spot (plume) basalts, referred to in Section 2.11, the mantle appears still to hold at least half of its early inventory of volatiles, and their weakening effect occurs at small concentrations and is little affected by further additions. The latent heat question is really asking: are we correctly calculating the very early heat loss at ocean ridges? As considered in the following section, the total heat lost by the oceanic lithosphere in its surface lifetime is about $2.9 \times 10^{14}\,\mathrm{J\,m^{-2}}$ and dwarfs the latent heat of any plausible depth of lava. Ridge lava poses an interesting question of detail on ridge behaviour but does not materially affect the thermal history of the mantle.

The core cooling history presents a completely different set of problems, a central one being the question of its radioactivity. This has been a subject of debate for several decades. Section 2.8 summarizes the chemical arguments and Stacey and Loper (2007) re-examine the physical argument. The case for radiogenic heat arises from the conductive heat loss and therefore depends on core conductivity (Section 19.6), which is calculated from the electrical conductivity. The preferred model adopted in Sections 21.4 and 22.7 has a small heat contribution by ^{40}K (0.2 terawatt at the present time) and Section 23.5 examines the implications for core cooling.

The extrapolation of the thermal regime of the Earth back to about 4.5 billion years, as presented in this chapter, is based on physical processes that are assumed to be smooth and continuous. The starting point is an Earth that has stabilized, following accretion and settling out of the core, with a fully solidified mantle and excess accretion energy lost to space. We reach some robust conclusions that hardly depend on assumptions. In particular, the thermal history is virtually the same, whether linear or non-linear rheology is assumed, and, although the cooling rate depends on what is assumed about the radioactive content of the Earth, the surface heat flux for the last four billion years hardly does so.

23.2 The rate of heat transfer to the oceans

The purpose of this section is to relate the ocean floor heat flux to the speed of the mantle convection that delivers this heat. The relationship allows us to substitute heat flux for the convective strain rate in the creep law (Eq. (10.27)) and so relate it to mantle temperature.

Approximately 3.4 km^2/year of new ocean lithosphere is produced at the ocean ridges and, of course, the same area disappears at subduction zones. With the total oceanic area of $3.1 \times 10^8\,\mathrm{km}^2$ the mean surface lifetime is 91 million years. Combining this with the average heat flux from ocean floors, 0.101 $\mathrm{W\,m^{-2}}$ (Pollack *et al.*, 1993), the average heat release by oceanic lithosphere in its surface lifetime is the product of these numbers, $2.9 \times 10^{14}\,\mathrm{J\,m^{-2}}$. The heat is transferred to the sea by a combination of thermal diffusion and hydrothermal circulation and observations indicate that hydrothermal cooling has more or less shut down by the time the lithosphere has reached an age of 50 or 70 million years, although we suggest that internal hydrothermal circulation may continue. The inference is that the sea floor cracks, through which the water circulates, become choked with sediment and this must be at least partly true, but there is another mechanism. As the lithosphere ages, so the depth from which the water must draw heat increases, reducing the temperature gradient driving the water circulation and increasing the viscous drag on its flow in the cracks. At the same time the diffusion of heat is reduced by the hydrothermal cooling of shallow layers, invalidating Eqs. (20.7) and (20.9). Having no theory of the overall cooling process, we assume a simple empirical relationship and adjust it to match the observations of heat flow and ocean depth discussed in Section 20.2.

The required relationship must give enhanced heat flux from young lithosphere, relative to the $t^{-1/2}$ variation in Eq. (20.7), but still allow significant heat from ageing lithosphere. We represent the observations by a simple power law for the average heat flux per unit area,

$$\dot{q} \propto t^{-a}. \tag{23.1}$$

This gives the integrated heat in the lithospheric life-time, τ,

$$Q \propto \tau^{1-a}. \tag{23.2}$$

The selected value of a is a compromise between conflicting requirements. We emphasize the stabilization of the depth curve for old lithosphere, which gives the integrated heat flux and is less dependent on details of the cooling model than is the heat flux itself. This requires $0.5 < a < 1.0$, and we choose $a = 2/3$, so that $(1 - a) = 1/3$ in Eq. (23.2). This is assumed in the calculations that follow, but we emphasize that it is not a theoretical result. It is a simple empirical fit to present day data, which we assume to be valid also in the distant past. The total heat, Q, takes time τ to pass into the sea, so the average rate is $\dot{Q} = Q/\tau \propto \tau^{-2/3}$. This is calibrated by the present average ocean floor heat flux, $\dot{Q}_0 = 0.101\,\mathrm{W\,m^{-2}}$, for an average lithospheric lifetime, $\tau_0 = 91 \times 10^6$ years (2.87×10^{15} s), that is

$$\dot{Q} \propto \tau^{-a} \tag{23.3}$$

or

$$\dot{Q} = \dot{Q}_0(\tau_0/\tau)^{2/3}. \tag{23.4}$$

We must relate τ to the speed of convection and for this it is necessary to know how the plate size varies. At the present time there is a wide range of both plate sizes and speeds. Carlson *et al.* (1983) reported a correlation, such that speed was roughly proportional to the square root of age at subduction, as might be expected if speed is proportional to shrinkage, and therefore to negative buoyancy, by the diffusive model. But such an argument cannot be extended to a similar relationship between average plate size and speed in the past because, with a variable mantle temperature and viscosity, there was no constant relationship between plate speed and driving stress. However, we can establish a general rule for average plate size by a series of substitutions, (i) to (iv), of proportionalities between the quantities involved. We consider the general case of arbitrary a in Eqs. (23.1) and (23.2).

(i) Start with $\tau^{-a} \propto \dot{Q}$ by Eq. (23.3).

(ii) Substitute for $\dot{Q}$ by using the thermodynamic result, from Chapter 22, that $\dot{Q} \propto \sigma\dot{\varepsilon}$ because the dissipation (stress × strain rate) is related to convected heat by a constant efficiency, independent of convective speed.

(iii) Substitute $\sigma \propto Q$, because the convective stress is proportional to the cumulative cooling of a slab, and then substitute for Q by Eq. (23.2).

(iv) Take strain rate $\dot{\varepsilon} \propto v$, the plate speed, and then write $v = L/\tau$, where the plate size, L, is identified as the distance from ridge source to subduction zone.

These give

$$\tau^{-a} \propto \dot{Q} \propto \sigma\dot{\varepsilon} \propto Q\,\dot{\varepsilon} \propto \tau^{1-a}\,\dot{\varepsilon} \propto \tau^{1-a}v \\ \propto \tau^{1-a}(L/\tau) \propto L\tau^{-a}. \tag{23.5}$$

We see that, by this line of reasoning, L is independent of τ and therefore of convective speed. Thus we can relate convective speed to the heat flux,

$$\dot{\varepsilon} \propto v \propto \tau^{-1} \propto \dot{Q}^{1/1a} \tag{23.6}$$

or

$$\dot{Q} \propto \dot{\varepsilon}^{a} = \dot{\varepsilon}^{2/3} \tag{23.7}$$

This result depends on the analysis, summarized by Eq. (23.5), leading to the conclusion that the mean plate size has not varied. It requires the mean plate speed to have been greater in the past and we can consider the implications, and the possibility that contradictory evidence will come to light. With fixed plate size, Eq. (23.3) or (23.7) gives speed proportional to $\dot{Q}^{3/2}$. By the cooling history in Section 23.4, 10^9 years ago this quantity would have been greater than at present by the factor 1.25. If we consider just the last 180 million year record of ocean floor magnetic stripes, the factor is 1.04. Plate speeds differ from one another now by much more than this and there is no prospect that paleomagnetism would be able to distinguish such variations in the average.

Equation (23.7) refers to the ocean floor heat flux per unit area. The total oceanic heat flux, $\dot{Q}_1$, is therefore related to the present value, $\dot{Q}_{1,0}$, by

$$\dot{Q}_1 = \dot{Q}_{1,0}(f/f_0)(\dot{\varepsilon}/\dot{\varepsilon}_0)^{2/3}, \tag{23.8}$$

where f is the fraction of the surface area of the Earth occupied by oceans and, from Pollack *et al.* (1993), we take the present values to be $f_0 = 0.606$ and $\dot{Q}_{1,0} = 3.10 \times 10^{13}$ W, being 0.101 $\mathrm{Wm^{-2}}$, less radiogenic heat of the ocean crust. Subscript 1 is introduced because there is a second component of the heat flux, through the continents.

The heat flux from the continents is largely due to crustal radioactivity, which must be discounted in calculating the mantle cooling. We estimate the total crustal radiogenic heat as 8.2×10^{12} W. But there is also a flux of heat through the continents from the mantle and in Section 20.3 this is estimated to be 0.025 $\mathrm{W\,m^{-2}}$. Thus we add to Eq. (23.8) a second component of the mantle heat loss, proportional to the fraction $(1 - f)$ of the surface area A that is occupied by continents,

$$\dot{Q}_2 = (1 - f)A \times 0.025\,\mathrm{W\,m^{-2}} \\ = 1.275 \times 10^{13}(1 - f)\,\mathrm{W}. \tag{23.9}$$

At the present time this is 11% of the total heat flux from the Earth and the fraction decreases backwards in time. For a reason considered in Section 9.3, we make the simple assumption that continental structure, in particular its thickness, remains essentially the same through time. This means that the area is proportional to the total mass of continental material that has accumulated and that the heat flux per unit area through it from the mantle has varied rather little, although its own radiogenic heat has decreased with time. This is a simplifying approximation, but it affects only a small component of the heat flux. The total heat lost by the mantle is the sum of Eqs. (23.8) and (23.9), less a steady component, $\dot{Q}_C = 3.5 \times 10^{12}$ W, which we attribute to the core and assume to be conveyed to the surface by plumes operating independently of the plate tectonic mechanism of mantle convection. Thus we write the mantle heat loss as

$$\dot{Q} = \dot{Q}_1 + \dot{Q}_2 - \dot{Q}_C. \tag{23.10}$$

The fraction f of the surface area occupied by oceans has decreased with time as continental material separated from the mantle. This is a consequence of convection and the rate has not been constant with time. We assume that it has been proportional to the rate of convective heat transport from the mantle, that is, to $\dot{Q}$ by Eq. (23.10). This means that the area of the continents at any time is proportional to the integrated heat flux since the origin of the Earth at $t = -T$. Thus we can write

$$(1-f)/(1-f_0) = (1-f)/0.394 = Q_{-T}^{-t}/Q_{-T}^{0}, \tag{23.11}$$

where Q_{-T}^{-t} is the integrated mantle heat over the period from $T = 4.5 \times 10^9$ years to t years ago and superscript zero means $t = 0$. But $Q_{-T}^{0} = Q_{-T}^{-t} + Q_{-t}^{0}$, so that Eq. (23.11) can be written in a more convenient form:

$$\frac{f}{f_0} = 1 + \left(\frac{1}{f_0} - 1\right)\frac{Q_{-t}^{0}}{Q_{-T}^{0}} = 1 + 0.65\frac{Q_{-t}^{0}}{Q_{-T}^{0}}. \tag{23.12}$$

The calculation is iterative because Q_{-T}^{0} is a product of the calculation and is repeatedly adjusted in seeking a solution. There is another role for the factor f. The ratio $(1-f)/(1-f_0)$ is the fraction of the present continental crust that has been emplaced at any time. Thus the fraction still in the mantle was

$$\frac{f - f_0}{1 - f_0} = \frac{Q_{-t}^{0}}{Q_{-T}^{0}}. \tag{23.13}$$

This time-dependent fraction of the radiogenic heat identified with the composition of the present crust is added to the mantle radiogenic heat. It means that we assume all of the radiogenic heat to have been in the mantle 4.5×10^9 years ago.

23.3 The heat balance equation and mantle rheology

The rate at which the mantle loses heat is the difference between the heat lost to the surface, $\dot{Q}(T_p)$, and radiogenic heating, $\dot{Q}_R(t)$. It is written as a heat balance equation,

$$\phi_m \frac{dT_p}{dt} = \dot{Q}_R(t) - \dot{Q}(T_p), \tag{23.14}$$

where $\phi_m = 7.4 \times 10^{27}$ J K^{-1} is the heat capacity, defined as the heat lost per degree fall in the potential temperature, T_p, as in Section 21.3. For this purpose the heat lost to the surface is only the heat lost by the mantle and excludes core heat, although the core heat is included in the surface heat flux calculation in Section 23.2 and subtracted in Eq. (23.10). Without any detailed knowledge of the two functions on the right-hand side of Eq. (23.14), we have a qualitative assessment of thermal history. $\dot{Q}_R(t)$ is a decreasing function of time, t, and $\dot{Q}(T_p)$ decreases as temperature falls, because it is controlled by convection, which slows up as the mantle stiffens. If we suppose the mantle to be in thermal balance, with $dT_p/dt = 0$, then in the moderately recent past $\dot{Q}_R$ would have been stronger and the mantle would have been heating up. Since, in that case, it would have been cooler, the convected heat loss, $\dot{Q}$, would have been smaller, reinforcing the temperature change. By this hypothesis the Earth started cool and has been heating up until now, making the present time unique, with a cooler Earth in both the past and the future. This is not plausible. The Earth started hot and, by Eq. (23.14), it is still cooling and will continue to do so.

The function $Q_R(t)$ is a sum of the exponential decays of the four thermally important isotopes

$$\dot{Q}_R = \dot{Q}_{R0}\,[f_1\exp(-\lambda_1 t) + f_2\exp(-\lambda_2 t) + f_3\exp(-\lambda_3 t) + f_4\exp(-\lambda_4 t)], \tag{23.15}$$

where f_1, f_2, f_3, f_4 are the fractional contributions to the present radiogenic heat, $\dot{Q}_R$, and t is time relative to the present, that is negative for past time. These fractions rely on an assessment of composition. We assume the Th/U ratio to be 3.7

Table 23.1 Fractions of present radiogenic heat attributed to four isotopes

Isotope	Decay constant λ (year^{-1})	Mantle heat fraction	Crustal heat fraction
1 ^{238}U	$1.551\,25 \times 10^{-10}$	0.4566	0.3892
2 ^{235}U	9.8485×10^{-10}	0.0197	0.0167
3 ^{232}Th	4.9475×10^{-11}	0.4763	0.4062
4 ^{40}K	5.544×10^{-10}	0.0474	0.1878

for both the mantle and crust, but that the K/U ratio is 2800 for the mantle but 13 000 for the crust. With these relative concentrations and the energies in Table 21.2 we obtain the fractional contributions in Table 23.1. We do not have a secure value for the present mantle radiogenic heat, $\dot{Q}_{R0}$. In the following section we adopt 20×10^{12} W as a preferred value but investigate the implications of alternatives. Equation (23.15) applies also to the crust, with the alternative fractions in Table 23.1 and $\dot{Q}_{R0} = 8.2 \times 10^{12}$ W in this case. The crustal radioactivity is progressively added to the mantle in the backwards integration, according to Eq. (23.13).

Now we consider the second term on the right-hand side of Eq. (23.14), that is the mantle heat loss. The present value, $\dot{Q}_0$, is the total rate of heat loss by the Earth, 44.2×10^{12} W, less 8.2×10^{12} W of crustal radiogenic heat and 3.5×10^{12} W of core heat, as discussed in Section 23.5, leaving $\dot{Q}_0 = 32.5 \times 10^{12}$ W (Table 21.4). Fortuitously, this almost coincides with the ocean floor heat flux, as reported by Pollack *et al.* (1993), with near cancellation of the core heat conveyed to the ocean floors and mantle heat conducted through the continents. With substitutions for $\dot{Q}_1$ and $\dot{Q}_2$ by Eq. (23.8) and (23.9) and *f/f*$_0$ by Eq. (23.12) or (23.13), Eq. (23.10) gives

$$\dot{Q}(T_P) = \dot{Q}_{10}\left(1 + 0.65\frac{Q^0_{-t}}{Q^0_{-T}}\right)\left(\frac{\dot{\varepsilon}}{\dot{\varepsilon}_0}\right)^{2/3} + \dot{Q}_{20}\left(1 - \frac{Q^0_{-t}}{Q^0_{-T}}\right) - \dot{Q}_C. \tag{23.16}$$

To use Eq. ((23.16)), we need another relationship between $\dot{Q}$ and $\dot{\varepsilon}$. This is obtained by combining the creep law (Eq. (10.27)) with proportionality (ii) in Eq. (23.5), that is $\dot{Q} \propto \sigma\dot{\varepsilon}$, rewritten as

$$\frac{\sigma}{\sigma_0} = \left(\frac{\dot{Q}}{\dot{Q}_0}\right) \Big/ \left(\frac{\dot{\varepsilon}}{\dot{\varepsilon}_0}\right). \tag{23.17}$$

In Eq. (10.27) B is constant and although μ varies with depth its variation with time (or temperature) is slight enough to neglect, so the equation can be rewritten relative to present conditions (subscripted zero),

$$\frac{\dot{\varepsilon}}{\dot{\varepsilon}_0} = \left(\frac{\sigma}{\sigma_0}\right)^n \exp\left[gT_M\left(\frac{1}{T_0} - \frac{1}{T}\right)\right], \tag{23.18}$$

where T is understood to be T_P and $T_0 = 1700$ K is its present value. Combined with Eq. (23.17) this gives

$$\left(\frac{\dot{\varepsilon}}{\dot{\varepsilon}_0}\right)^{n+1} = \left(\frac{\dot{Q}}{\dot{Q}_0}\right)^n \exp\left[gT_M\left(\frac{1}{T_0} - \frac{1}{T}\right)\right], \tag{23.19}$$

which we use to substitute for $(\dot{\varepsilon}/\dot{\varepsilon}_0)$ in Eq. (23.16) by rewriting as

$$\left(\frac{\dot{\varepsilon}}{\dot{\varepsilon}_0}\right)^{2/3} = \left(\frac{\dot{Q}}{\dot{Q}_0}\right)^{2n/3(n+1)} \exp\left[\frac{2gT_M}{3(n+1)}\left(\frac{1}{T_0} - \frac{1}{T}\right)\right], \tag{23.20}$$

so that

$$\dot{Q} = \dot{Q}_{10}\left(1 + 0.65\frac{Q^0_{-t}}{Q^0_{-T}}\right)\left(\frac{\dot{Q}}{\dot{Q}_0}\right)^{2n/3(n+1)} \exp\left[\frac{2gT_M}{3(n+1)}\left(\frac{1}{T_0} - \frac{1}{T}\right)\right] + \dot{Q}_{20}\left(1 + \frac{Q^0_{-t}}{Q^0_{-T}}\right) - \dot{Q}_C. \tag{23.21}$$

This is the equation for $\dot{Q}(T_P)$ required in Eq. (23.14). Together with the relationship for mantle radiogenic heat, with component 1 for the present mantle composition and 2 for the present crust, which is progressively added to the mantle in the backward integration

$$\dot{Q}_R = \dot{Q}_{R1} + \frac{Q^0_{-t}}{Q^0_{-T}}\dot{Q}_{R2}, \tag{23.22}$$

it allows Eq. (23.4) to be integrated numerically. For each of a range of values of $\dot{Q}_{R0}$, and $n = 1$ or $n = 3$, Eq. (23.14) is integrated to find (by iteration) values of the parameter g in

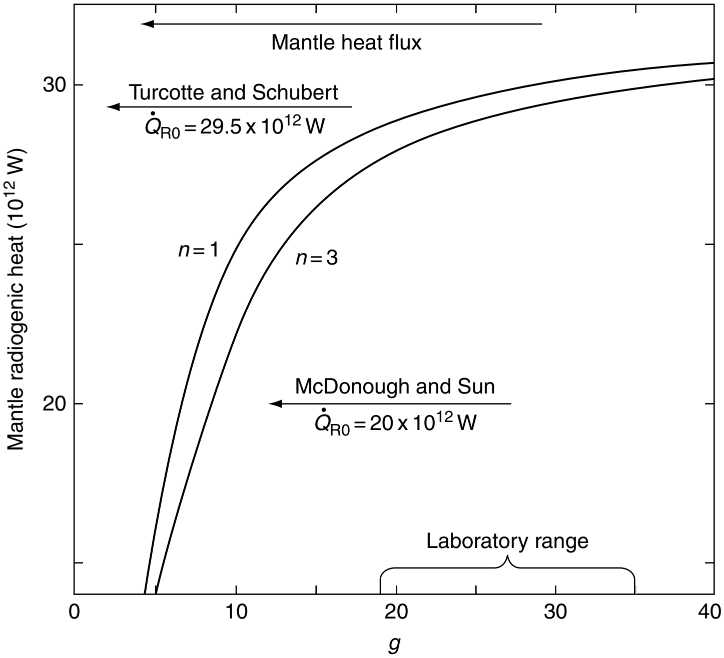


FIGURE 23.1 Mutual dependence of the mantle radiogenic heat, $\dot{Q}_{R0}$, and the rheological parameter, g, for two values of n, $n = 1$ for linear (Newtonian) rheology and $n = 3$ for non-linear rheology. Values of $\dot{Q}_{R0}$ are indicated for compositions by McDonough and Sun (1995) and Turcotte and Schubert (2002).

Eq. (23.21) that give an extrapolation back to $T_P = 2399$ K at $t = -4.5 \times 10^9$ years. This is the assumed starting temperature, constrained by matching it to the mantle solidus at the core–mantle boundary, where it is assumed to coincide with the core temperature at that time. The results are plotted in Fig. 23.1. We need to note that we are using a creep equation that applies to material at a specified temperature with an identifiable melting point, but both T and T_M have wide ranges in the mantle. We use the potential temperature, T_p, to represent the mantle as a whole, and no error or difficulty would arise if the ratio of temperature to melting point were the same everywhere but that is not the case. The ratio T/T_M decreases with depth over most of the mantle range and this is reflected in the viscosity variation. So, we appeal to the concept of averaged or notional values of both T and T_M, recognizing that this compromises the meaning (and numerical value) of the parameter g in Eqs. (10.27) and (23.21). But an exponential dependence of rheology on temperature, as in Eq. (10.27), is unavoidable, so, although we cannot impose theoretical values of quantities such as g or T_M, we can reach some important general conclusions that are not sensitive to them.

23.4 Thermal history of the mantle

Integration of Eq. (23.14) requires an assumption about starting conditions, that is, the notional mantle temperature 4.5×10^9 years ago. The only thing known for sure about conditions at that time is that they were changing very rapidly and simple extrapolations back from the present cannot convey any information about the accretion process. We are considering the evolution of the Earth from the stage when it had stabilized sufficiently for cooling to be controlled by convection in a solidified mantle, with no significant continuing accretion, loss of volatiles to space or extended separation of the core. The starting point is assumed to be early enough for there to be no thermal boundary layer at the base of the mantle, attributable to greater cooling of the mantle than of the core. Thus, referred to the core–mantle boundary temperature, core and mantle thermal histories start from the same point. A different assumption is possible. We could suppose that the core retained more of the accretion energy, and started much hotter than the mantle, so that there would have been a strong thermal boundary layer at the base of the mantle from the beginning. However, in our thermal history model the mantle starts at its solidus temperature, so if the core were much hotter it would have melted the mantle to a considerable depth, allowing it to convect so rapidly that the excess core heat would have been quickly removed, bringing the Earth to the starting point that we assume, with the mantle at its solidus temperature.

Core cooling (Section 23.5) is constrained by the requirement that the inner core has existed for most of the life of the Earth. By Eq. (21.14), this means a decrease in core–mantle boundary (CMB) temperature no more than 200 K. Since we argue that the mantle has cooled by about 1000 K more than this, it is useful to have some corroboration of the slow CMB cooling. In a consideration of melting points of mantle minerals, Boehler (2000) concluded that 'The extrapolated solidus of the lower mantle intersects the geotherm at the core–mantle boundary . . .'. He pointed out that this is consistent with the identification of ultra low velocity zones (ULVZs) at the base of the mantle as pockets of partial melt. The enormous accretion energy (Table 21.1) ensured that the Earth started hot, but any molten stage would have lost heat very rapidly, leaving the mantle at its solidus temperature, which is therefore the starting point for convective cooling controlled by solid state creep (Eq. 10.27). Thus Boehler's conclusion invites the supposition that the core–mantle boundary has not cooled at all. But a growing inner core, so important to the dynamo (Section 22.7), disallows this. Mao *et al.* (2006) presented an alternative explanation for ULVZs. They observed that the post-perovskite phase (ppv), to which perovskite is converted at the base of the mantle, readily absorbs iron and that iron-rich ppv has acoustic velocities compatible with ULVZs, without appealing to partial melting. We suppose that heterogeneities in the mantle, especially in the D″ layer at its base, give a range of solidus temperatures and that a better average value to assume for the starting point of the mantle cooling calculation is 100 K to 200 K higher than the present CMB temperature. Taking this to be 3931 K and extrapolating adiabatically to $P = 0$, we have an initial mantle potential temperature, $T_{p0} = 2399$ K. This is a notional temperature for the application of Eqs. (10.27) and (23.14) to the mantle as a whole. It is also the notional melting point for the purpose of these equations, that is we assume the mantle to start at its melting point. In Eq. (23.21) we see that this is not a critical assumption because T_M appears in a single adjustable constant with the rheological parameters g and n, which are not well constrained.

We treat the parameters n and g in Eqs. (10.27) and (23.21) and $\dot{Q}_{R0}$ in Eq. (23.15) as adjustable, but they are related in the sense that if one is fixed then integration of the model to the assumed starting conditions at $t = -4.5 \times 10^9$ years constrains the others. Geochemical arguments, largely derived from observed radioactivity in meteorites and in the crust, favour $\dot{Q}_{R0} \approx 20 \times 10^{12}$ W. This is the value obtained from the pyrolite model of McDonough and Sun (1995). We adopt it for our preferred model, although we argue for a different K/U ratio (Table 23.1) so the assumption is rather arbitrary. We note the much higher estimate of the mantle content of radioactive isotopes quoted by Turcotte and Schubert (2002, Table 4.2, p. 137), 29.5×10^{12} W, and consider both this case and an intermediate value, 24×10^{12} W, as alternatives to the preferred model.

The mutual constraint on values of $\dot{Q}_{R0}$, n and g is illustrated by Fig. 23.1. The first thing to notice is that the value of n makes little difference to the relationship between $\dot{Q}_{R0}$ and g. For both $n = 1$ and $n = 3$, if g is within the range of laboratory observations, then $\dot{Q}_{R0}$ must have a high value, quite close to the present mantle heat loss, $\dot{Q}_0 = 32.5 \times 10^{12}$ W, making the present cooling rate very slow with very rapid early cooling. Much smaller values of g are required if the radiogenic heat comes within the range of the geochemical estimates. Neither of the estimates of radiogenic heat marked on Fig. 23.1 can be regarded as secure, which is the reason for considering also the intermediate value, 24×10^{12} W.

FIGURE 23.2 The geothermal flux for three values of mantle radiogenic heat. This is the total, with crust and core heat added to the mantle heat flux. Differences between curves for alternative values of *n* are not noticeable on this scale. The rate of heat loss is not very dependent on parameters of the cooling theory for most of the life of the Earth. The major difference occurs very early.

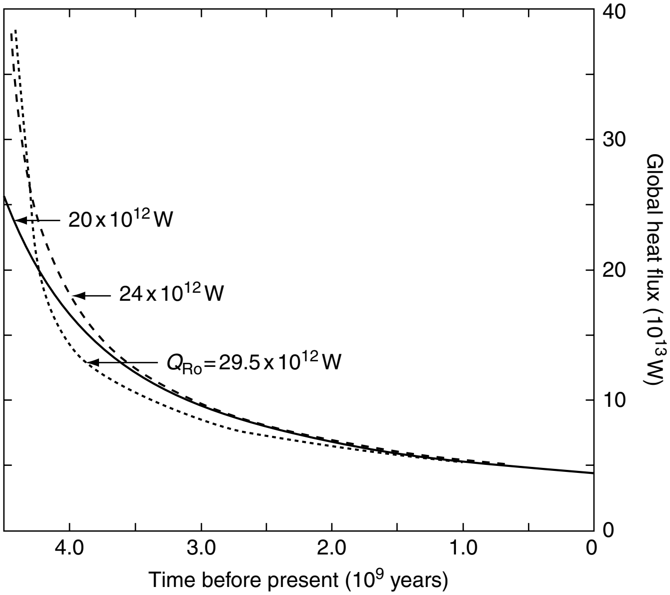


FIGURE 23.3 Mantle temperature variation as a function of time for three values of radiogenic heat. Rapid early cooling is a feature of all models, more so for those with stronger radioactive heating.

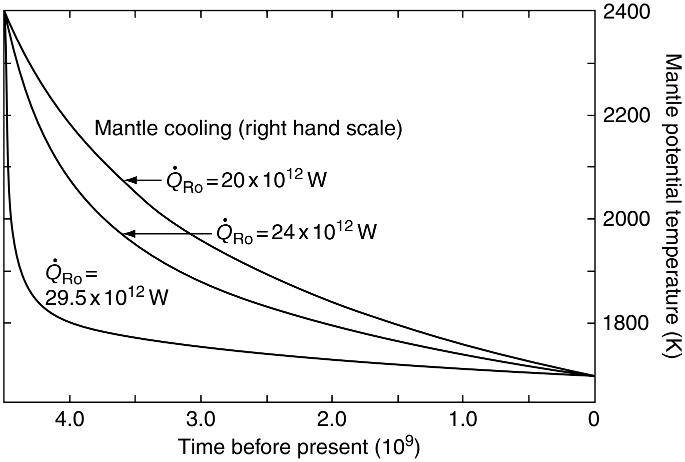


Figure 23.2 is a plot of the variation with time of the total heat flux from the Earth for the three models. On this scale the differences between plots for $n = 1$ and $n = 3$ do not show and for most of the life of the Earth the heat loss curves are not very different. The big difference occurs only in the first 100 million years. This is a consequence of the constraint imposed by matching the models to the present rate of heat loss, 44.2×10^{12} W. The temperature curves (Fig. 23.3) differ according to the fraction of the heat loss that is made up by radioactivity, but in this case also the $n = 1$ and $n = 3$ curves differ too little to show separately. For the 29.5×10^{12} W model the present mantle cooling rate is 10 K/billion years, corresponding to 16 K/billion years at the core–mantle boundary, less than the estimated core cooling rate. This appears unlikely because the mantle must have cooled faster than the core to allow any core cooling at all, and so far as we understand the mechanism this trend would not have reversed. We consider the

24×10^{12} W model to represent an upper bound on mantle radiogenic heat. For this, and the preferred 20×10^{12} W model, the cooling curves of Fig. 23.3 are not very different and offer nothing that would distinguish them observationally.

We note from Fig. 23.1 that the preferred model requires a value of g differing by a factor of two from the laboratory range, and accept that the laboratory values are not appropriate to the application of Eq. (10.27) to the mantle as a whole. This is not surprising, as the model assumes a uniform homologous temperature, T/T_M, although it varies with depth. A more complete model might allow for this, but we see that it would have little influence on the conclusions. In Eq. (23.21) T_M is coupled with g and n in a composite parameter, $2gT_M/3(n+1)$, and does not enter the calculations independently. Uncertainties in g and n are certainly greater than the uncertainty in T_M and, as Fig. 23.1 illustrates, the essential conclusions of the cooling model hardly depend on the value of n. They are, therefore, not very dependent on the naivety of the assumption about T_M. Although details are uncertain, we see in these models common features that we regard as secure. Mantle convection governed by an exponential dependence of viscosity on temperature, as in Eq. (10.27), gives a plausible thermal history that is almost independent of whether the rheology is linear or non-linear. The rate of terrestrial heat loss depends significantly on the assumed radiogenic heat only in the first 10^8 years or so, when validity of the model can be questioned anyway. For all models early cooling is fastest, giving a substantial thermal boundary layer at the base of the mantle for most of the life of the Earth. Mantle radiogenic heat is not well determined, but the preferred estimate, 20×10^{12} W, is unlikely to err by more than 4×10^{12} W. If we specify both the present rate of heat loss and radiogenic heat then the thermal history follows almost independently of assumptions about physical properties. The essential physics is embodied in the two obvious principles, strongly temperature-dependent rheology and decaying radiogenic heat, with little ambiguity in the temperature variation and virtually none in the terrestrial heat flux.

Table 23.2 Numerical details of the thermal history model with 20×10^{12} W of mantle radiogenic heat

Present radiogenic heat	
Mantle	20×10^{12} W
Crust	8.2×10^{12} W
Core	0.2×10^{12} W
Total	28.4×10^{12} W
Present rate of heat loss	
Mantle	32.5×10^{12} W
Crust	8.2×10^{12} W
Core	3.5×10^{12} W
Total	44.2×10^{12} W
Radiogenic heat in 4.5×10^9 years	7.6×10^{30} J
Total heat loss in 4.5×10^9 years	13.4×10^{30} J
Residual (stored) heat	13.3×10^{30} J
Radiogenic heat now to $t = \infty$	10.9×10^{30} J

Table 23.2 summarizes numerical details of the preferred model, with 20×10^{12} W of mantle radiogenic heat and a core heat loss of 3.5×10^{12} W, as discussed in Section 23.5. This extends some of the discussion in Chapter 21, in which it is pointed out that radiogenic heat accounts for little more than half of the total heat loss in the life of the Earth (Table 21.1). The fraction of the Earth's heat loss that is attributed to radioactivity is known as the Urey ratio. Its present value, for our model, is 0.64. Over the life of the Earth the ratio is 0.57, so it has varied rather little. The longevity of ^{232}Th ensures that in the distant future it dominates the radiogenic heat and it may be surprising that the radiogenic heat still to be released exceeds the heat released so far in the life of the Earth. Internal heat will maintain tectonic activity until conduction suffices to carry the mantle heat flux and that state will not be reached for more than 10^{10} years.

23.5 Cooling history of the core

The cooling of the core is controlled by the properties of the mantle and the behaviour of the thermal boundary layer at its base (D″). Thermal processes in the core are consequences, not causes, of the heat loss. This was the basis of an argument by Stacey and Loper (1984) that the heat flux from the core is stabilized to a more or less steady value by a balance between competing effects in the boundary layer. Heat is removed from the core by a convective process that we identify with plumes of hot, low viscosity material that flows rapidly up narrow channels in the mantle. But all convective processes are slowed by the cooling and stiffening of the mantle. The competing effect is the relatively faster cooling of the mantle than the core, which increases the temperature increment across the boundary layer, causing an increase in the heat flux into it which compensates for the general stiffening. Although we cannot expect this compensation to be more than approximate, it is essential to understanding what happens in the core. The point is that it is the heat flux that is stabilized and not the rate of change in temperature. We now consider how these are related.

Table 21.5 distinguishes several contributions to core energy that vary differently with time. We ignore, for the moment, the contributions by radioactivity and precession and divide the first five entries in the table into heat $Q_1 = 20.23 \times 10^{28}$ J, resulting from cooling at a rate proportional to $-\mathrm{d}T/\mathrm{d}t$, and $Q_2 = 13.91 \times 10^{28}$ J, arising from latent heat and compositional separation, which contribute at a rate proportional to $\mathrm{d}V/\mathrm{d}t$, where V is the inner core volume. $\mathrm{d}V/\mathrm{d}t$ and $\mathrm{d}T/\mathrm{d}t$ are related by Eq. (22.41). Since $r \propto V^{1/3}$, we can rewrite this as

$$V^{-1/3}\mathrm{d}V/\mathrm{d}t \propto \mathrm{d}T/\mathrm{d}t, \tag{23.23}$$

which we can integrate, without regard to the time dependence, from $V=0$ at $T=T_0$ to V at temperature T. Substituting present values, V_p and T_p to eliminate the proportionality constant,

$$\left(\frac{V}{V_\mathrm{p}}\right)^{2/3} = \frac{T_0 - T}{T_0 - T_\mathrm{p}}. \tag{23.24}$$

By differentiating we relate the time dependences of V and T

$$\frac{1}{V_\mathrm{p}}\frac{\mathrm{d}V}{\mathrm{d}t} = \frac{3}{2}\left(\frac{T_0 - T}{T_0 - T_\mathrm{p}}\right)^{1/2}\left(\frac{-\mathrm{d}T/\mathrm{d}t}{T_0 - T_\mathrm{p}}\right). \tag{23.25}$$

The extremes of this relationship are $\mathrm{d}V/\mathrm{d}t \to 0$ at ($T = T_0$, $V = 0$) and $\mathrm{d}V/\mathrm{d}t \to (3/2)V_\mathrm{p}/(T_0 - T_\mathrm{p})$ at ($T = T_\mathrm{p}$, $V = V_\mathrm{p}$).

We identify V/V_p with the fraction of heat Q_2 released, so $\mathrm{d}Q_2/\mathrm{d}t$ varies from zero to $(3/2)(Q_2/\Delta T)(-\mathrm{d}T/\mathrm{d}t)$, where $\Delta T = 98.4$ K is the temperature drop accompanying inner core formation. The variation in Q_1 is linear in temperature, so $\mathrm{d}Q_1/\mathrm{d}t = (Q_1/\Delta T)(-\mathrm{d}T/\mathrm{d}t)$. Thus the total rate of heat release at any stage is

$$\dot{Q} = \left[\frac{3}{2}\left(\frac{T_0 - T}{T_0 - T_\mathrm{p}}\right)^{1/2}\frac{Q_2}{\Delta T} + \frac{Q_1}{\Delta T}\right]\left(\frac{-\mathrm{d}T}{\mathrm{d}t}\right). \tag{23.26}$$

From the values of Q_1 and Q_2 above, the fraction contributed by the Q_2 term varies from zero to 0.508 as the inner core grows from nothing to its present size. If, as we suppose, the mantle control keeps $\dot{Q}$ approximately constant then $\mathrm{d}T/\mathrm{d}t$ decreases by a factor 2 during inner core growth. But this does not compensate for the low thermodynamic efficiency of the Q_1 sources (0.15), compared with the Q_2 sources (0.50). Maintenance of constant dynamo power requires a decrease in the core-to-mantle heat flux with time to offset the increasing thermodynamic efficiency with the greater contribution by inner core growth. By the cooling calculation presented here, inner core formation began early, at least 3.5 Ga ago, and has become progressively more important as its increasing contribution to dynamo power compensated for a decreasing cooling rate.

The core-to-mantle heat flux implied by this calculation, about 3.5 terawatt, has some observational support because it corresponds to the plume heat flux inferred from Sleep's (1990) estimate of the buoyancy of hot spot 'swells'. This identifies the hot spots with plumes originating at the base of the mantle and carrying up the core heat. Most other estimates of the core–mantle heat flux are higher, typically 10 tW. Even with 1 to 2 tW of radiogenic heat, as often assumed, this would mean core cooling faster than mantle cooling for all of the mantle cooling models in the previous section (see Problem 23.3). But the mantle must have cooled much more than the core to leave a thermal boundary layer at its base and it would be difficult to devise a thermal history with this conflict in relative cooling rates.

When we introduce radiogenic heat to the argument, we have more early heat, but if $\dot{Q}$ is constrained by the mantle control, the added heat merely slows the early cooling rate by substituting a heat source of even lower thermodynamic efficiency. Precession introduces some early energy of high efficiency but, in Section 24.7, we conclude that it is not important at the present time and would have been only a minor contribution, even early in the life of the Earth.

Now we take a closer look at Eqs. (23.24) and (23.26) to see how drastically we must modify the constant $\dot{Q}$ assumption. Suppose that 50% of the heat $(Q_1 + Q_2)$ is used, as we might assume to be the case two billion years ago. The values of V and T are obtained by writing

$$Q_1\left(\frac{T_0 - T}{T_0 - T_P}\right) + Q_2\left(\frac{V}{V_P}\right) = Q_1\left(\frac{T_0 - T}{T_0 - T_P}\right) + Q_2\left(\frac{T_0 - T}{T_0 - T_P}\right)^{3/2} = \frac{1}{2}(Q_1 + Q_2). \quad (23.27)$$

With the values of Q_1 and Q_2 above, we have a numerical solution $(T_0 - T)/(T_0 - T_p) = 0.5576$. The corresponding fractional volume is $V/V_p = 0.4164$, that is the inner core is less than half formed although the required temperature drop has passed the half-way point. Now we can use $(T_0 - T)/(T_0 - T_p) = 0.7467$ in Eq. (23.26) to find the fraction of $\dot{Q}$ contributed by the Q_2 term at that stage of inner core development. The result is 0.435, compared with the present value, 0.508.

A conclusion from Eq. (23.27) is that the Q_1 term was dominant early in the development of the inner core but that for the second half of its development, perhaps the last two billion years, the ratio of the Q_1 and Q_2 contributions varied too little to question the constant $\dot{Q}$ assumption. In Section 25.4 we mention reports of paleomagnetic observations suggesting that the geomagnetic field was systematically weaker 2 to 2.5 billion years ago than in more recent times. This is consistent with a slightly smaller contribution to core energy by the inner core at that time. However, when the inner core was much smaller, there is no alternative to a stronger core-to-mantle heat flux, whether attributed to more rapid cooling or to radioactivity. The constant $\dot{Q}$ assumption is not compatible with the observation that the geomagnetic field existed 3.5 billion years ago. But the inference that it appears satisfactory for the last two billion years is consistent with our argument that radiogenic heat has a minor role.

We have little observational control on theories of the early state of the core, but conclude that it was never more than 200 K hotter than at present. There is even a suggestion, reviewed by Sumita and Yoshida (2003), that the core was initially stably stratified. The idea is that, as the Earth accreted, the core formed by liquid iron sinking through the proto-mantle, with its solutes in local chemical equilibrium with mantle minerals. But the solubility of oxygen in liquid iron increases with pressure, so that the early accumulating inner part of the core would have had less oxygen than the later part because it percolated through the mantle at lower pressure. The resulting stable stratification would not have broken down until development of the inner core began to release excess oxygen into the liquid. It is important not to contemplate any theory that allows doubt about the viability of the dynamo. If this stable stratification occurred it could not have survived for a billion years, and requires a very early start to inner core development, but that is compatible with the cooling model presented here.

We see that, with the core conductivity that we assume, it is marginally possible to avoid the need to assume radiogenic heat in the core. In view of the uncertainty in the conductivity, we must consider the case for radioactivity in the core to remain open for discussion. This is not the only doubt. The calculations all assume that the rejected light solute from inner core solidification mixes uniformly into the outer core. If we assume instead that it all finds its way to the top of the core then the gravitational energy is 1.84 times as much as for uniform mixing. Such a dramatic enhancement of compositional power would greatly strengthen the case against core radioactivity, but there is no serious evidence for it. A consequence would be a gravitationally stable light layer at the top of the core, with a mass deficit Δm (Eq. (22.30)), or perhaps a large fraction of it, relative to uniform composition. It would appear as a surface layer of thickness t (m) with a density deficit $\Delta\rho = (2.8 \times 10^7/t)\ \mathrm{kg\,m^{-3}}$ relative to the deeper core. If the volume of the layer is comparable to the inner core volume then its thickness is $t \approx 50$ km and $\Delta\rho \approx 560\ \mathrm{kg\,m^{-3}}$. The existence of a stably stratified layer with a thickness of this order has been suggested on the basis of geomagnetic studies (Whaler, 1980; Braginsky, 1993, 1999) but doubted (Fearn and Loper, 1981). If even a small fraction of the light concentrate reached the core–mantle boundary it would establish a stable layer, so the possibility must be allowed, but the fraction would have to be a large one if it is to have a significant influence on core energetics. There is no seismological evidence requiring such a layer, but it would not be easily obtained. No P-waves have their deepest penetration in the outermost part of the core and observations would need to use a weak S-to-P conversion at the core–mantle boundary.

24

The geomagnetic field

24.1 Preamble

The study of geomagnetism has a longer history than other branches of geophysics, partly because of its use as an aid to navigation. In the earliest times, the spontaneous alignment of magnetized iron oxide (lodestone) was believed to be due to an extra-terrestrial influence. Properties of spherical lodestones were described in a letter, written in 1269 by Petrius Peregrinus (Pierre de Maricourt of Picardy in France). He introduced the word 'poles' in connection with magnets because their influence was believed to be derived from the celestial poles. Recognition of the similarity of the magnetic fields of lodestones to the field of the Earth appears to have awaited the work of Robert Norman and William Gilbert of England in the sixteenth century. A review of their evidence that the Earth is a great magnet, with its axis approximately north–south, was presented, in Latin, in Gilbert's book *De Magnete*, published in 1600. An English translation by P. F. Mottelay appeared in 1893 and has been reprinted (Dover Publications, 1958) as an early milestone in scientific literature.

Magnetic declination, the difference between the direction of the field, as indicated by mariners' compasses, and true (geographic) north, was indicated on navigational maps by the mid-sixteenth century, but was attributed to a misalignment of magnetic and geographic axes and not to a flaw in Gilbert's concept of a dipole field. There were also observations of a progressive change in declination in London that were not mentioned by Gilbert, although it would be surprising if he were unaware of them. This is a manifestation of what we now know as the geomagnetic secular variation, which was first reported in 1634 by H. Gellibrand. Edmund Halley, of comet fame, made a detailed study of magnetic declination and its secular variation, initially over the Atlantic Ocean and then more widely. By 1700 he recognized not only that there are significant departures from a dipole field, but that features of the field showed a generally westward drift. He attributed this to inner concentric shells, rotating slightly more slowly than the outer shell of the Earth and carrying with them embedded magnets, so anticipating by more than two centuries modern ideas about differential rotation within the core. The westward drift has featured prominently in geomagnetic studies and is discussed in Section 24.6.

Following Halley's work, there was only one notable development before the twentieth century. In 1838, C. F. Gauss applied a development of potential theory (the basis of spherical harmonic analysis – Appendix C) to confirm beyond doubt the inference of Gilbert and Halley that the field is of internal origin. By that time the connection between rapid disturbances of the field and aurorae was well documented and the only observed natural events that appeared to be related to geomagnetism occurred outside the Earth. Thus, Gauss's analysis may have stalled a drift back to Peregrinus's notion of an externally generated field. Gauss must have been aware of experiments on electromagnetic induction by M. Faraday, published six years earlier, but he did not make the connection and there is no evidence of any suggestion that the geomagnetic field was driven by deep electric currents until about 1900. Even then it was not readily accepted, and studies of geomagnetism were completely dominated by relationships between magnetic disturbances and extra-terrestrial effects, solar activity and aurorae.

The self-exciting dynamo mechanism that we now accept as the cause of the geomagnetic field faced conceptual difficulties that were gradually removed by four developments.

(i) In 1908, G. Hale, director of the Mount Wilson astronomical observatory, reported the splitting of spectral lines in the radiation from sunspots that could be caused only by very strong magnetic fields (the Zeeman effect).

(ii) Hale's observation aroused speculation on the mechanism for spontaneous generation of sunspot fields by motion of a fluid conductor, prompting Larmor (1919) to suggest that this could explain not only a solar field but also the terrestrial field.

(iii) H. Alfven developed a theory of cosmic-scale magnetic fields controlled by and controlling the motion of tenuous plasma. He originated the frozen flux concept of a magnetic field carried around and deformed by a fluid conductor, showing that amplification was not only possible but inevitable with suitable turbulent motion.

(iv) Seismology established that the Earth has a dense fluid core. The cosmic abundance of iron, apparent in meteorites and in the solar atmosphere, pointed to a liquid iron core.

In the 1940s, W.M. Elsasser began putting these ideas together in the first serious study of the magnetohydrodynamic dynamo mechanism. He was soon followed by E.C. Bullard, who was quick to recognize the significance of what Elsasser had started. But rival hypotheses were slow to die. Theoretical physicists, led by Einstein, were groping for evidence of a connection between gravity and electromagnetism, postulating a fundamental relationship between the rotations and magnetic fields of large bodies. This particular hypothesis received its fatal blow only with very sensitive observations on rotating spheres, which induced no magnetic fields (Blackett, 1952), and observations that the strength of the geomagnetic field generally increased, rather than decreased, with depth in mines.

In spite of early statements, such as one in 1902 by L.A. Bauer (cited by Parkinson, 1983, p. 108) that the field was due 'doubtless to a system of electric currents embedded deep within the interior of the Earth and connected in some way with the earth's rotation', dynamo theory faced strong and influential opposition. Especially negative was S. Chapman who, as late as 1940 (Chapman and Bartels, 1940, p 704), reasserted an earlier statement by A. Schuster: 'the difficulties which stand in the way of basing terrestrial magnetism on electric currents inside the earth are insurmountable'. Ideas about geomagnetism have evolved over many years. Complementary perspectives on its history are presented by Elsasser (1978, pp. 225–230), Parkinson (1983) and Merrill *et al.* (1996).

Our observations of the geomagnetic field are limited by the fact that it is generated in the core, which is little more than half the radius of the Earth. Small-scale features are invisible, not just because they are diminished by distance, but because they are obscured by magnetization in the Earth's crust. Core-generated features of the field smaller than about 1500 km are concealed. Our view of the field is restricted also by the electrical conductivity of the mantle, which is low compared with that of the core but high enough to attenuate magnetic fluctuations with periods shorter than a year or so. Estimates of the conductivity profile of the mantle have been based on the assumption that it varies with radius, but not laterally. Although this gives a global view it can be no more than an approximation and, with respect to the lowermost mantle, may be seriously wrong. A long-standing observation that some features of the field appear to be stationary while others drift (Yukutake and Tachinaka, 1969) is consistent with the idea that the 'standing' features are held in place by high conductivity patches in the mantle. In view of the heterogeneity of D″, which is presumed to be both compositional and thermal, this is the most plausible location for such patches. Although this is conjectural, it introduces a complication to the inference of core motions from extrapolations of the surface field to the core–mantle boundary. For a recent discussion of inferred core motion see Eymin and Hulot (2005).

Unmaintained electric currents in the core would decay by ohmic dissipation with a time constant of order 10^4 years. The general features of a regeneration mechanism, involving a combination of convection and rotation, are understood (Section 24.5). Both thermal and compositional convection are presumed to occur, with the compositional effect dominant at the present time. Core energetics are discussed in Section 22.7. The possibility of a contribution by precessional torques is also canvassed (Malkus, 1963, 1989; Vanyo, 1991), but, by our estimate of the angular difference between core and mantle angular momentum axes (Section 7.5), is only a minor effect. There is no precise estimate of the energy demand by the dynamo, but we can appeal to the principle that whatever mechanism occurs most easily is the one that dominates. 10^{11} W may well suffice, but a consideration of nutational coupling suggests three times as much and, in calculating core energy, we assume a dynamo power requirement of 3×10^{11} W, as in Fig. 21.4.

Although the secular variation of the geomagnetic field is precisely observed, the rate of change is slow and the historical record of direct observations gives only limited insight. To see how the field behaves on a longer time scale we turn to paleomagnetism, the record of the field in ancient times, preserved in the magnetizations of rocks and archeological specimens. When we look at the field on a geological time scale, we see a new range of phenomena, that could not even be guessed at from the historical record. This is the subject of Chapter 25, but the discoveries of paleomagnetism, especially field reversals and the axial dipole principle, have had a profound effect on our understanding of the field, as discussed in this chapter.

24.2 The pattern of the field

As William Gilbert noticed, the magnetic field of the Earth is similar to that of a magnetized sphere, or to a small but powerful bar magnet at the centre. Such a field is termed a dipole field because it could, in principle, be produced by a pair of magnetic poles of equal strengths but opposite signs a small distance apart. The magnetic moment or dipole moment, m, is then envisaged as the product of pole strength and separation. Since magnetic moments are produced by circulating currents, the equivalent definition in terms of a current loop is the product of current i and loop area A,

$$m = iA. \tag{24.1}$$

Although the Earth's field is predominantly dipolar, non-dipole components contribute about 20% of its strength at the surface. We refer to a best-fitting dipole and need to be specific about what this means. The dipole most closely fitting the observed field is slightly off centre, but if we refer to the best-fitting geocentric dipole then its moment (in 2005) is

$$m_{\mathrm{Earth}} = 7.768 \times 10^{22}\,\mathrm{A\,m^2}. \tag{24.2}$$

The axis of this dipole is inclined to the geographic or rotational axis by 10°. If we discount the equatorial component of this moment and consider just the geocentric axial dipole then its moment is $7.644 \times 10^{22}\,\mathrm{A\,m^2}$.

The field of a dipole is conveniently represented in terms of the scalar magnetic potential, V_m, which may be differentiated to obtain any component of the field,

$$V_m = \frac{\mathbf{m} \cdot \mathbf{r}}{4\pi r^3} = \frac{m \cos\theta}{4\pi r^2}, \tag{24.3}$$

where θ is the angle between the dipole axis and the radius vector $\mathbf{r}$ from the dipole to the point considered. These equations assume the dimensions of the current loop to be negligible compared with r. The field is

$$\mathrm{B} = -\mu_0 \text{ grad } V_m. \qquad (24.4)$$

The horizontal (circumferential) and vertical (radial) components are

$$B_\theta = -\frac{\mu_0}{r}\frac{\partial V_m}{\partial \theta} = \frac{\mu_0}{4\pi}\cdot\frac{m}{r^3}\sin\theta, \qquad (24.5)$$

$$B_r = -\mu_0\frac{\partial V_m}{\partial r} = \frac{\mu_0}{4\pi}\cdot\frac{2m}{r^3}\cos\theta. \qquad (24.6)$$

Applying these equations to the Earth's surface, taken to be a sphere of radius a, and noting that they refer to the magnetic axis, not the geographic axis, the horizontal and vertical components of the field, in conventional notation, are

$$H = -B_\theta(r=a) = -B_0\sin\theta, \qquad (24.7)$$

$$Z = -B_r(r=a) = -2B_0\cos\theta, \qquad (24.8)$$

where

$$B_0 = \frac{\mu_0}{4\pi}\frac{m_{\mathrm{Earth}}}{a^3} = 3.004\times10^{-5}\,\mathrm{T} = 0.3004\,\mathrm{Gauss} \qquad (24.9)$$

is the strength on the magnetic equator of the best-fitting geocentric dipole field. By choosing the coordinate axis to be the axis of the dipole we avoid consideration of a latitudinal component of the field, since $\partial V_m/\partial\lambda = 0$. In the SI system of units, used here, $\mu_0 = 4\pi\times10^{-7}\,\mathrm{H\,m^{-1}}$ is the permeability of free space. In the cgs system $B_0 = 0.3004$ Gauss and is the same as $H_0 = 0.3004$ Oersted, with $\mu_0 = 1$ Gauss/Oersted. The total field strength at the Earth's surface $(r=a)$ is

$$B = \left(B_\theta^2 + B_r^2\right)^{1/2} = B_0\left(1+3\cos^2\theta\right)^{1/2}. \qquad (24.10)$$

Its inclination to the horizontal is I, where

$$\tan I = B_r/B_\theta = 2\cot\theta = 2\tan\phi \qquad (24.11)$$

and $\phi = (90° - \theta)$ is the magnetic latitude. Equation (24.11) is used to calculate paleopole positions from the dip angles of magnetic remanence in rocks. It is also the differential equation for a magnetic line of force because

$$\tan I = \frac{\mathrm{d}r}{r\mathrm{d}\theta} = 2\cot\theta, \qquad (24.12)$$

which integrates to

$$\frac{r}{a} = \frac{\sin^2\theta}{\sin^2\theta_a}, \qquad (24.13)$$

where θ_a is the magnetic co-latitude at which the field line crosses radius a, generally taken to be on the magnetic equator.

By subtracting the dipole field from the observed field, we are left with the non-dipole field. Its features, as they were in 1945, are clearly displayed in Fig. 24.1, which is one of the maps produced by Bullard *et al.* (1950), who sought clues to the origin of the field by studying the behaviour of the non-dipole field. For the International Geomagnetic Reference Field (IGRF2005, Table 24.1), the rms strength of the non-dipole field over the Earth's surface, 1.11×10^{-5} T, is a quarter of the rms strength of the dipole field, $\sqrt{2}B_0 = 4.248\times10^{-5}\,\mathrm{T}$, but, as discussed below, the ratio is quite different at core level.

The separation of the dipole and non-dipole fields is accomplished by spherical harmonic analysis. Equation (24.3) is the first term of an infinite sum with the form of Eq. (C.11). As noted in Appendix C, by convention in geomagnetism, tesseral and sectoral harmonics are normalized to have the same rms values over a spherical surface as the zonal harmonics of the same degree, but they are not fully normalized in the sense of the p_l^m harmonics in Appendix C. The system used in geomagnetism is referred to as Schmidt normalization after A. Schmidt, who introduced it. Harmonic coefficients of the field, g_l^m and h_l^m, obtained with this convention are the Gauss coefficients of the field. For the complete internal field, that is omitting terms that represent a field of external origin,

$$V_m = \frac{a}{\mu_0}\sum_{l=1}^{\infty}\left(\frac{a}{r}\right)^{l+1}\sum_{m=0}^{l}\left(g_l^m\cos m\lambda + h_l^m\sin m\lambda\right) \times \mathrm{P}_l^m(\cos\theta). \qquad (24.14)$$

This is so written that the coefficients, g and h, have the dimensions of field; they are normally given in nanoTesla (nT), that is, 10^{-9} T or 10^{-5} Gauss, often referred to as gammas in older literature. Coefficients up to $(l,m) = (8,8)$ are given in Table 24.1a. The use of satellite data for global magnetic field modelling, discussed by Olsen

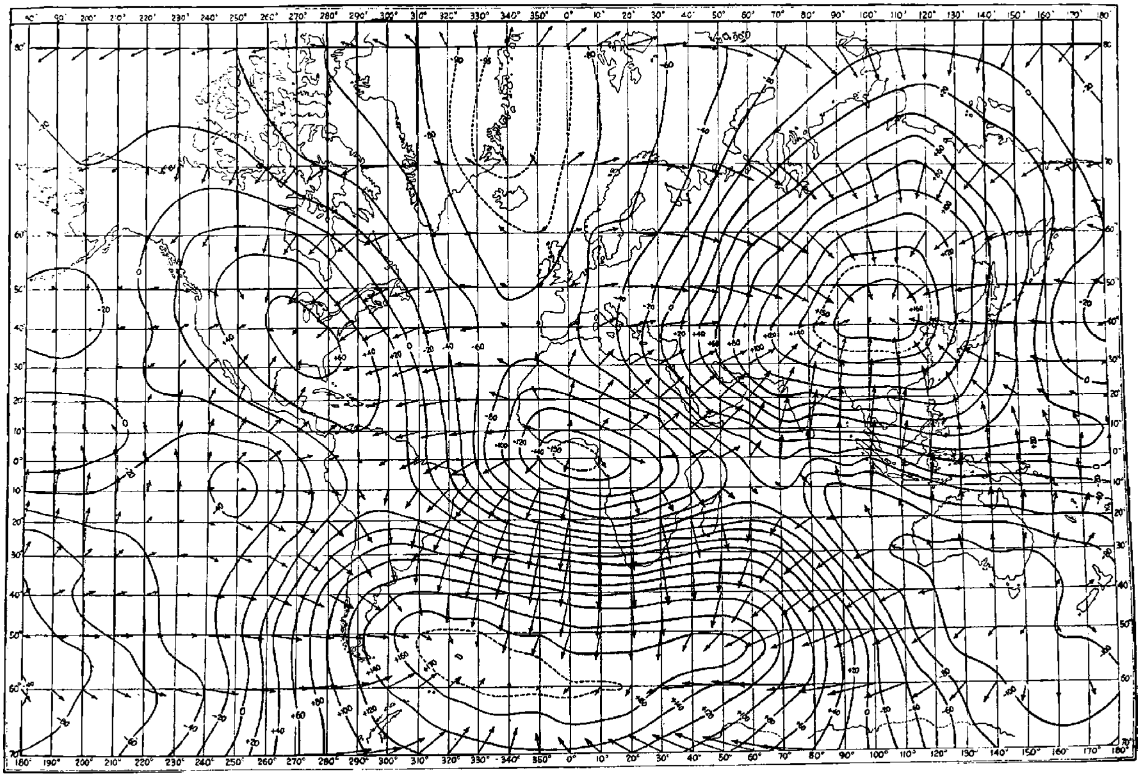


FIGURE 24.1 The non-dipole field for 1945. Contours give the vertical component in intervals of 2μ T and arrows represent the horizontal component. Reproduced, by permission, from Bullard *et al.* (1950).

Table 24.1a Spherical harmonic coefficients of the International Geomagnetic Reference Field (IGRF) 2005 to degree and order 8, 8. For each (l,m) the coefficients are g_l^m followed by h_l^m in nanoTesla, as defined by Eq. (24.14), referred to a surface of radius 6371.2 km

m *l*	0	1	2	3	4	5	6	7	8
1	−29556.8	−1671.8							
		5080.0							
2	−2340.5	3047.0	1656.9						
		−2594.9	−516.7						
3	1335.7	−2305.3	1246.8	674.4					
		−200.4	269.3	−524.5					
4	919.8	798.2	211.5	−379.5	100.2				
		281.4	−255.8	145.7	−304.7				
5	−227.6	354.4	208.8	−136.6	−168.3	−14.1			
		42.7	179.8	−123.0	−19.5	103.6			
6	72.9	69.6	76.6	−151.1	−15.0	14.7	−86.4		
		−20.2	54.7	63.2	−63.4	0.0	50.3		
7	79.8	−74.4	−1.4	38.6	12.3	9.4	5.5	2.0	
		−61.4	−22.5	6.9	25.4	10.9	−26.4	−4.8	
8	24.8	7.7	−11.4	−6.8	−18.0	10.0	9.4	−11.4	−5.0
		11.2	−21.0	9.7	−19.8	16.1	7.7	−12.8	−0.1

Table 24.1b Secular variation of the IGRF 2005. These are the rates of change (nT/year) of the coefficients in Table 24.1a

m	0	1	2	3	4	5	6	7	8
l									
1	8.8	10.8							
		−21.3							
2	−15.0	−6.9	−1.0						
		−23.3	−14.0						
3	−0.3	−3.1	−0.9	−6.8					
		5.4	−6.5	−2.0					
4	−2.5	2.8	−7.1	5.9	−3.2				
		2.0	1.8	5.6	0.0				
5	−2.6	0.4	−3.0	−1.2	0.2	−0.6			
		0.1	1.8	2.0	4.5	−1.0			
6	−0.8	0.2	−0.2	2.1	−2.1	−0.4	1.3		
		−0.4	−1.9	−0.4	−0.4	−0.2	0.9		
7	−0.4	0.0	−0.2	1.1	0.6	0.4	−0.5	0.9	
		0.8	0.4	0.1	0.2	−0.9	−0.3	0.3	
8	−0.2	0.2	−0.2	0.2	−0.2	0.2	0.5	−0.7	0.5
		−0.2	0.2	0.2	0.4	0.2	−0.3	0.5	0.4

(2002), leads particularly directly to the spherical harmonic representation in Eq. (24.14).

There is no $l = 0$ term in Eq. (24.14) because that would correspond to a magnetic monopole. If monopoles exist they certainly do not contribute to geomagnetism. The $l = 1$ terms represent the dipole. For this general representation of the field, the axis is the geographic axis but the co-latitude, θ, is the geocentric value, which differs from the geographic co-latitude by a small angle due to the ellipticity of the Earth (Fig. 6.2). The coefficient g_1^0 is the axial component of the dipole field and the equatorial components are given by g_1^1 (through the Greenwich meridian) and h_1^1. The strength of the dipole field on the magnetic equator is

$$B_0 = \left[(g_1^0)^2 + (g_1^1)^2 + (h_1^1)^2 \right]^{1/2}, \tag{24.15}$$

as given by Eq. (24.9). The angle between the magnetic and geographic axes is η, where

$$\tan \eta = \left[(g_1^1)^2 + (h_1^1)^2 \right]^{1/2} / |g_1^0| = 0.192, \tag{24.16}$$

giving $\eta = 10.26°$. Higher degree terms in Eq. (24.14) represent quadrupole, octupole and higher multipole components of the field.

Field components in the northward (X), eastward (Y) and downward (Z) directions are derivatives of V_m:

$$X = \frac{\mu_0}{r} \frac{\partial V_m}{\partial \theta}, \tag{24.17}$$

$$Y = \frac{-\mu_0}{r \sin \theta} \frac{\partial V_m}{\partial \lambda}, \tag{24.18}$$

$$Z = \mu_0 \frac{\partial V_m}{\partial r}, \tag{24.19}$$

at geocentric colatitude θ and longitude λ. All of these components are directly observable, whereas V_m is not. If all three components are measured, then there is some redundancy in calculating the g and h coefficients, but the additional information is needed if the field is not assumed to be entirely internal in origin. Then another set of coefficients, corresponding to C' and S' in Eq. (C.11) and representing

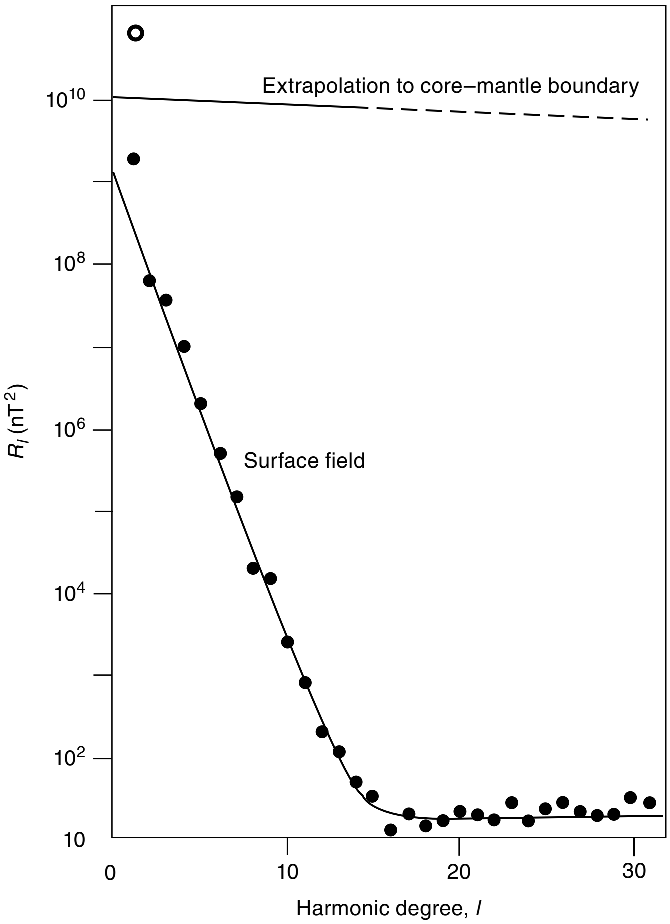


FIGURE 24.2 Mean square field strength, R_l, (Eq. 24.23), for each harmonic degree, l, plotted as a function of l. Solid circles are data points from Cain *et al.* (1989). The open circle and the upper line are downward continuations to the core–mantle boundary of the dipole and the trend of harmonics $l=2$ to 14. Harmonics up to $l=63$ were included in the original analysis but are increasingly affected by noise.

external sources, must be separated from g and h. This is discussed in Section 24.3 in connection with the separation of fields induced within the Earth by external, disturbing fields.

The normalization of spherical harmonics is discussed in Appendix C. When this is applied to the vector magnetic field, to obtain the strength or intensity, and averaged over the surface of a sphere, the rms field strength at the Earth's surface, due to any harmonic term represented by g_l^m or h_l^m, is (Lowes, 1966)

$$\left(B_l^m\right)_{\text{rms}} = (l+1)^{1/2}\left(g_l^m, h_l^m\right). \tag{24.20}$$

Spherical harmonics are orthogonal functions, so that in summing terms we add their squares, and the mean square surface field due to all harmonic terms of degree l is

$$R_l = (l+1)\sum_{m=0}^{l}\left[\left(g_l^m\right)^2+\left(h_l^m\right)^2\right]. \tag{24.21}$$

We can regard this as the sum of all terms representing wavelengths ($2\pi a/l$), so that a plot of R_l vs l (Fig. 24.2) is a power spectrum of the spatial variation in field strength. This plot shows two distinct ranges. At low harmonic degrees the core field is dominant but falls off rapidly with l, whereas for $l > 14$ it is masked by the field due to crustal magnetization, which has an almost white spectrum (R_l nearly independent of l). Cain *et al.* (1989) fitted the data to an equation with two terms, one representing each source,

$$R_l = 9.66 \times 10^8 (0.286)^l + 19.1(0.996)^l \text{ nT}^2. \tag{24.22}$$

The form of Eq. (24.14) allows the surface field to be extrapolated downwards to its source, a mathematical process known as downward continuation in exploration geophysics, in which it is used to estimate the depths and shapes of sources of magnetic and gravitational anomalies. Since the potential of an lth degree harmonic falls off with radius as $r^{-(l+1)}$, the field components (Eqs. (24.17) to (24.19)) fall off as $r^{-(l+2)}$ and, therefore, the square of the field and the sum of squares of all degree l components, that is R_l, falls off as $r^{-(2l+4)}$. Thus, the mean square field at radius r, for all terms of degree l, is

$$R_l(r) = R_l(a) \cdot (a/r)^{2l+4}, \tag{24.23}$$

where the value at the surface (radius a) is obtained from the coefficients in Table 24.1a. The upper line in Fig. 24.2 is the downward continuation to the core–mantle boundary by application of Eq. (24.23) to the first term in Eq. (24.22). At that depth the core-generated field has a spectrum that is pink (almost white, but still slightly red). Similarly, the crustal field, observed at the surface, is spectrally slightly pink, although that is not obvious in Fig. 24.2.

The field within a source region cannot be estimated by downward continuation without further assumptions because it can be represented by a potential, such as Eq. (24.14), only outside the source volume. However, the spectrum at the surface of a source can be a useful indicator of its dimensions. If we consider a source of infinitesimal thickness at a single depth, then from the spectrum at any higher level we can determine the depth if the source spectrum is known, or determine the source spectrum if the depth is known. If spectral contributions from different levels in a distributed source are independent, their effect is simply additive at the surface. For a spectrally uniform source occupying a depth range that is small compared with the radius r, the spectrum at the surface is very similar to that for a spectrally similar but very thin source located in the middle of the depth range. So, although downward continuation into a source region is formally disallowed, by downward continuing to the mid-depth of a thin source we can obtain a satisfactory measure of the source spectrum.

The near whiteness of both crustal and core fields at source level invites the inference that the sources are in fact white and thin. This is readily justified in the case of the crustal field. By Eq. (24.22), the mid-depth of a spectrally white crustal field would be at a radius $(r/a) = \sqrt{0.996}$, 13 km below the surface. This coincides with the median depth of the magnetized crust, because material below 20 to 25 km is at temperatures above the Curie points of magnetic minerals. Hence, the crustal magnetization has a white spatial spectrum. The significance of a similar extrapolation for the core field (the first term of Eq. (24.22)) is less obvious. The spectral calculations of Cain *et al.* (1989), used in this equation, give a median depth for a spectrally white source 80 ± 47 km below the core surface. Several authors (e.g. Langel and Estes, 1982) have given similar estimates, in spite of the long extrapolation from the Earth's surface.

We take the view that the nearly white spatial spectrum of the field at core level is a physically significant feature with a fundamental cause, although there is not, to our knowledge, an explanation in dynamo theory. Also, it is not clear that the principle applies to other planets and, in any case, it is necessary to exclude the dipole and probably the quadrupole terms, which do not fit the spectrum. But we suppose that the white spectrum describes the smaller-scale features that indicate core turbulence. We consider a simple model that is consistent with the observations and is useful to the discussion of the secular variation of the field in Section 24.3. Adopting the 80 km half-depth estimate, we assume that a characteristic depth of about 160 km in the surface of the core is a physically real feature. The current pattern is spectrally white but the field at greater depths is unobservable. By the frozen flux principle (Sections 24.3 and 24.5), the observable field moves with the outer layer of the core, which screens the deeper parts from view. The observed spectrum is characteristic of the field in the surface layer and gives no direct clue to the deeper pattern. The 160 km scale has a straightforward explanation as the size of the smallest magnetically controlled vortex in the core motion. We therefore identify it with the upper bound of the spatial spectrum of the field at core level, corresponding to harmonic degree $l = 136$, and model the core surface field by current loops of all radii down to this limit.

Although the dipole field ($l = 1$) is not as dominant at core level as it is at the surface, its strength is clearly above the trend of the higher harmonics (Fig. 24.2, open circle). The difference is more than a statistical effect, indicating that the dipole field has a special role in the geodynamo, so that we give it separate consideration in calculating the total rms field strength at core level. The quadrupole ($l = 2$) field is lower than the trend and this, too, probably has a physical cause in the dynamo, but it is not separately identified in the numbers that follow. The relationship between these field components is considered in a more general way in Sections 25.3 and 25.4, using evidence of a fundamental distinction between harmonics that are symmetric or antisymmetric about the equator. Omitting the dipole term, the downward-continued core field at the core–mantle boundary has a spectrum represented by the upper line in Fig. 24.2 and given by

$$R_l(\mathrm{CMB}) = 1.085 \times 10^{10}(0.959)^l\ \mathrm{nT}^2. \qquad (24.24)$$

This may be used to calculate the total rms field strength, $\sqrt{(\sum R_l)}$, for any range of harmonics. With an independent calculation for the dipole field from the $l = 1$ coefficients in Table 24.1, the sums of all field components of differently selected groups of harmonics (added in quadrature) are

$$\begin{aligned} B_{\mathrm{rms}}(\mathrm{CMB}) &= 2.61 \times 10^5\ \mathrm{nT\ for}\ l = 1, \\ &\quad 4.90 \times 10^5\ \mathrm{nT\ for}\ l = 2\ \mathrm{to}\ 136, \\ &\quad 5.55 \times 10^5\ \mathrm{nT\ for}\ l = 1\ \mathrm{to}\ 136. \end{aligned}$$

By this estimate the poloidal field strength at core level is about 5.5×10^5 nT (5.5 Gauss). If we consider the spectrum to be white 100 km or so inside the core then a higher value, 11.8×10^5 nT, applies. This number is derived from observations of the poloidal field, but it cannot be the total field in the core. Dynamo theory requires, in addition to the observed poloidal field (a field with poles), toroidal fields that are confined to the source region and are not observable in principle. The simplest toroidal field would be found in the material of a ring on which a current-carrying coil is uniformly wound. With care in winding the coil in layers progressing both backwards and forwards, so that there is no current component along the toroidal ring, the field is confined to it, forming a closed loop with no poles. From coupling of the inner core to nutations, the strength of the radial field penetrating it was estimated by Mathews *et al.* (2002) to be 7.2×10^6 nT. We interpret this as the total field strength within the core, including the toroidal field, because the inner core is a slightly better conductor than the outer core (Section 24.4) and can support the currents responsible for the toroidal field. On this basis, the toroidal field is about five times as strong as the poloidal field (see Section 24.7).

24.3 The secular variation and the electrical conductivity of the mantle

The slow variation in the geomagnetic field, driven by the changing pattern of motion in the core, is referred to as the secular variation. These changes occur on a time scale that extends from 1 year or so upwards, into the range of the variations caused by solar disturbances, which include an 11-year period related to the sunspot cycle. However, most of the extra-terrestrial effects are more rapid than any observable changes in the core field and no serious confusion arises.

As mentioned in Section 24.1, the historical record is too short to display some of the geomagnetic phenomena that have been recognized from the study of paleomagnetism. However, for the past few hundred years, and especially the last 50 years, we have a far more detailed picture of the behaviour of the field than could ever be expected for the remote past. This section is concerned primarily with the interpretation of the historical observations. Barton (1989) has given an overall review and a detailed analysis of the changing pattern of the field at core level

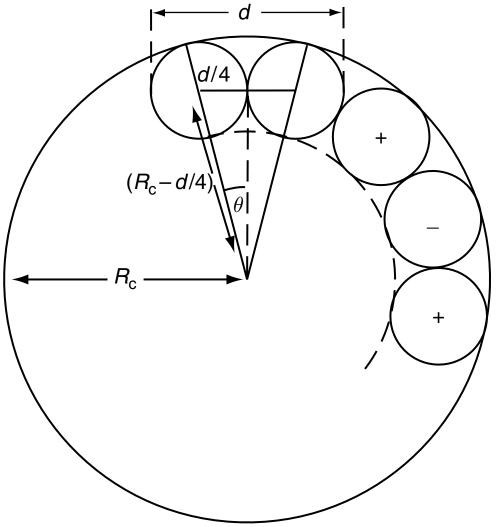


FIGURE 24.3 Geometry used to calculate the relationship between the size, *d*, and number, *l*, of current loops that can be fitted into the perimeter of the core (Eq. 24.29). The current loops are toroids, each represented by a pair of small circles.

is reported by Bloxham *et al.* (1989). Yukutake (1989) has summarized the relevant theories.

The time-variations of the spherical harmonic coefficients of the field, $\dot{g}$ and $\dot{h}$ (Table 24.1b), are tabulated only up to degree 8, that is for features with surface wavelengths of 5000 km or greater. Uncertainties in higher degree terms are too great for them to be meaningful. The corresponding wavelength limit at core level is 2700 km. It is evident from the table that even for $5 \leq l \leq 8$ the values of $\dot{g}$ and $\dot{h}$ are very small and are not precisely determined. Nevertheless, they suffice to give an idea of the relative rates of change of features of different sizes. A quantity analogous to R_l (Eq. (24.22)), but using the rates of change of the harmonic coefficients, is

$$Q_l = (l+1)\sum_{m=0}^{l}\left[\left(\dot{g}_l^m\right)^2+\left(\dot{h}_l^m\right)^2\right], \tag{24.25}$$

and we can define a reorganization time τ_l for the degree l components of the field,

$$\tau_l = (R_l/Q_l)^{1/2}. \tag{24.26}$$

We expect the reorganization times, τ_l, to be systematically related to the dimensions of current loops in the core that would produce the different harmonic features of the field.

Consider the geometry of Fig. 24.3, which represents a series of circular current loops, each of cross-sectional radius $d/4$, fitted into the core, radius R_c. We have

$$\sin\theta = \frac{d/4}{R_c - d/4}, \tag{24.27}$$

and if there are l current loops, that is l pairs of small circles in a complete circumference, then

$$4\theta = 2\pi/l \tag{24.28}$$

and the variation of loop diameter with harmonic degree, l, is

$$d = 4R_c[1 + 1/\sin(\pi/2l)]^{-1}. \tag{24.29}$$

For current loops of a particular shape and conductivity, the electromagnetic relaxation times vary as the square of the linear dimensions (Eq. (24.43)), so we might expect the degree reorganization times (Eq. (24.26)) to vary as

$$\tau_l \propto d^2. \tag{24.30}$$

Figure 24.4 is a plot of τ_l vs d, both on logarithmic scales, permitting the reasonableness of Eq. (24.30) to be judged by eye. The $l = 1$ (dipole) term stands above the trend and $l = 2$ below it. This appears to have fundamental significance and to be related to the dynamo mechanism, with magnetic energy fed from $l = 2$ to $l = 1$. Indeed, the trend for all the odd l harmonics is systematically above that for the even l terms. McFadden *et al.* (1988) model the secular variation in a way that distinguishes dipole and quadrupole 'families' of harmonics by their symmetry about the equator, and in Section 25.3 we refer to paleomagnetic evidence for this distinction. The non-dipole field is not simply fluctuating noise superimposed on the dipole but an essential feature of the dynamo.

With respect to the approximate validity of Eq. (24.30), ignoring the odd/even difference, we can take a linear regression of $\log\tau$ vs $\log d$ to compare the gradient with the value 2 suggested by electromagnetic relaxation. Using all eight values the gradient is 2.5 ± 0.5 (one standard deviation). If the dipole ($l = 1$) is omitted, then the gradient is 1.7 ± 0.4, and if both $l = 1$ and $l = 2$

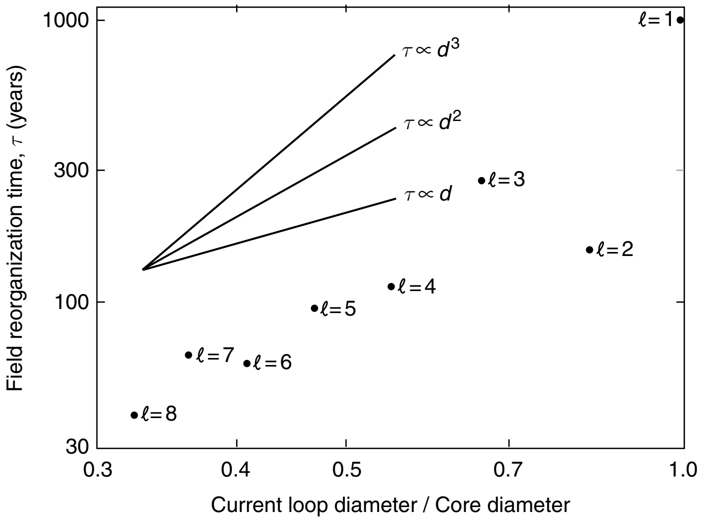


FIGURE 24.4 A plot of field reorganization time, τ_l, for harmonics of degree l (Eq. (24.26)), against the diameters of current loops in the outer core that would be responsible for these harmonics of the field (Eq. (24.29)). The field data used are those in Table 24.1. Numbers against the data points give the harmonic degrees.

terms are omitted the gradient is 2.4 ± 0.3. But the notion of a simple relaxation process cannot be valid. The values of τ are smaller by a factor 10 than the relaxation times for loops of the sizes modelled. The changes in field structure are not controlled by magnetic diffusion (or, equivalently, ohmic dissipation) but indicate the rate of rearrangement of the field by core motion. This is the frozen flux principle, by which a moving conductor carries the field with it. For this reason we refer to τ as a 'reorganization time' and not as a relaxation time.

A more detailed illustration of the frozen flux principle is presented by Bloxham *et al.* (1989), who deduced the pattern of secular variation at the core–mantle boundary from the observed surface field at a series of dates spanning the historical record. They followed particular features of the field in time, using as the definition of their boundaries the surrounding null-flux contours where the vertical component of the field is zero. These boundaries move about, but if the frozen flux principle applies then the total flux within a boundary remains constant. Bloxham *et al.* (1989) concluded that, although some diffusion is observable, it is a secondary effect. The frozen flux condition gives a valid representation of the observed secular variation. A finer scale of unresolved features must be superimposed but does not affect this conclusion. With the frozen flux condition the secular variation can be used to infer the motion of the core surface. Eymin and Hulot (2005) applied satellite data to this problem, concluding that some strong vortices were apparent, but that unseen fine-scale features of the field have important effects.

In Section 24.1 we mention that a westward drift has been a prominent feature of the secular variation for the duration of the historical record. However, when we examine archeomagnetic and paleomagnetic data spanning several thousand years a different picture emerges. A paleomagnetic perspective on secular variation and the westward drift is presented in Section 25.3. Before extensive paleomagnetic data became available and theories were constrained only by the historical record, a westward drift of features of the field by about 0.2 °/year, as found by Bullard *et al.* (1950), was generally supposed to be a permanent feature. This has appeared paradoxical because the frozen flux principle locks the field lines to the outermost part of the core and electromagnetic coupling would bring it to rotational equilibrium with the mantle in a decade or so. An attempt to explain the westward drift in terms of propagating hydromagnetic waves, first suggested by Hide (1966) remains of interest, whatever the duration of the drift. Braginsky (1991) emphasized the role in dynamo theory of travelling disturbances subject to magnetic, Archimedean (buoyancy) and Coriolis pressures and known by their acronym, MAC waves. The assumption is that they propagate along the toroidal magnetic field, B_T, in the core, and that this is much stronger than the observed poloidal field. In the approximation that the Archimedean pressure is neglected, the wave is known as a magneto-geostrophic mode and, for wavelength λ, has a phase speed

$$V_{\mathrm{MAC}} \approx \pi B_T^2/\mu_0\rho\omega\lambda, \tag{24.31}$$

where $\rho \approx 10^4\,\mathrm{kg\ m^{-3}}$ is the local density, $\omega = 7.292 \times 10^{-5}\,\mathrm{s^{-1}}$ is the Earth's rotational speed, and the characteristic dimension $\lambda/2$ corresponds to the features that are seen to drift, that is at least 4×10^6 m (4000 km). A wave with wavelength-dependent speed given by Eq. (24.31) is strongly dispersive. Applying Eq. (16.49), the group speed is

$$U_{\mathrm{MAC}} = 2V_{\mathrm{MAC}} \approx 2\pi B_T^2/\mu_0\rho\omega\lambda, \tag{24.32}$$

which is the speed of observable features. For a drift speed of 0.2°/year or $4 \times 10^{-4}\,\mathrm{m\,s^{-1}}$ at the core surface, $B_T \approx 0.02$ T (200 Gauss), that is 30 to 40 times as strong as the poloidal field. This explanation faces two difficulties. No strong dispersion is observed – the Bullard *et al.* (1950) analysis concluded that different harmonic components of the field drifted at the same rate. Also, the energy requirement of a dynamo maintaining a 0.02 T field is implausibly high. So, we seek another explanation for the persistence of the westward drift in the historical record. The only obvious possibility appears to be gravitational coupling of the inner core to the deep mantle (Section 24.6).

As we mention in Section 24.1, electrical conduction in the mantle influences our observations of the field. For the purpose of this discussion the mantle is solid and stationary, with only a passive role as the seat of induced electric currents, and it has a much lower conductivity than the core. The effect is essentially the same as the frozen flux principle, as applied to the core, in that the induced currents oppose motion of the field relative to the material of the conductor. The fixity of the mantle means that it resists changes to the field in it. This is Lenz's law of electromagnetic induction. In the case of an oscillatory field, the frequency of oscillation determines how effectively it penetrates the conductor. This is the skin effect. The amplitude of oscillation is reduced by a factor $1/e$ at a depth termed the skin depth, which is a function of frequency and conductivity. There is also a phase delay. For a plane wave, of angular frequency ω, entering a semi-infinite medium of conductivity σ, the variation with depth z of its amplitude and phase are given by

$$B = B_0 e^{-\alpha z}\sin(\omega t - \alpha z), \tag{24.33}$$

where

$$\alpha = 1/z_0 = (\mu_0\sigma\omega/2)^{1/2} \tag{24.34}$$

and z_0 is the skin depth.

The trends in the data plotted in Figs. 24.2 and 24.4 give no indication that the more rapid components of the secular variation (larger l or smaller τ) are noticeably attenuated by mantle conduction. The conductivity is not high enough to influence the observation of changes occurring over periods of 30 years or more. Evidence that changes occurring in a period of about 1 year can penetrate the mantle arose from an event in 1969–70, when there was a widespread, probably global, impulsive change in the secular variation, although possibly not synchronous everywhere. It is referred to as the 1969 geomagnetic jerk. It is best displayed in plots of time derivatives of the eastward component of the field (Y), because this is least affected by external disturbances. At many observatories there was a sharp change in $\mathrm{d}Y/\mathrm{d}t$, suggesting a discontinuity in $\mathrm{d}^2Y/\mathrm{d}t^2$. 'Sharp' in this context means that it occurred within about a year. Noise of external origin precludes any possibility of seeing more rapid effects if they were to occur. The jerk is surprising, both because it imposes a bound on mantle conductivity that is lower than previously supposed and because it is difficult to understand how a global change in core motions could occur so rapidly. It was some years before the possibility of an external influence was securely discounted, but the internal origin is not now in doubt (Courtillot and LeMouël, 1984). There is some evidence of earlier jerks, although less well documented.

As we note below, the mantle conductivity increases sharply at the 660 km phase boundary and so we concentrate attention on the transmission of a jerk through the more highly conducting lower mantle, depth range $\Delta z \approx 2200$ km. Consider the attenuation in amplitude, $|B|$, of a field component oscillating with angular frequency $\omega = \pi/\text{year} = 10^{-7}\,\text{s}^{-1}$ to represent the spectrum of the jerk, as it propagates upwards, ignoring its phase. By Eq. (24.33), over the range dz the variation is

$$\frac{\mathrm{d}|B|}{|B|} = -\alpha \mathrm{d}z = -(\mu_0 \sigma \omega/2)^{1/2} \mathrm{d}z, \qquad (24.35)$$

and integrating over a total depth range Δz,

$$\begin{aligned} \ln(|B|/B_0) &= -\sqrt{\mu_0 \omega/2} \int_0^{\Delta z} \sigma^{1/2} \mathrm{d}z \\ &= -\sqrt{\mu_0 \omega/2} \langle \sigma^{1/2} \rangle \Delta z. \qquad (24.36) \end{aligned}$$

If we take $|B|/B_0 = 0.5$, that is a 50% attenuation of the signal, then $\langle \sigma^{1/2} \rangle = 1.78\,\text{S}^{1/2}\,\text{m}^{-1/2}$, that is an average lower mantle conductivity of $3.2\,\text{S}\,\text{m}^{-1}$. Although this is a simplified, plane wave approximation to a spherical situation and assumes a uniform conductivity, it gives a reasonable order of magnitude estimate of the average lower mantle conductivity.

A complementary approach to mantle conductivity is to observe the effect of electromagnetic signals originating in upper atmospheric disturbances. Surface field variations are superpositions of external (inducing) and internal (induced) field variations. Being internal, the induced field can be represented by a potential with the spherical harmonic form of Eq. (24.14) or the terms with unprimed coefficients in Eq. (C.11) of Appendix C. The inducing (external) fields must be represented by the primed coefficients C' and S' in Eq. (C.11). The difference appears in the radial dependences. The potential itself is not observable, but its derivatives, the field components north (X), east (Y) and downwards (Z) are all recorded. Differentiating Eq. (C.11), ignoring the summation, that is considering only the zonal harmonic l (with $m = 0$), and then substituting $r = a$ for observations on the surface,

$$\begin{aligned} X &= -(1/r)(\partial V/\partial \theta) \\ &= -(1/a^2)(C_{lm} + C'_{lm})(\partial \mathrm{P}_{lm}(\cos\theta)/\partial \theta), \qquad (24.37) \end{aligned}$$

$$\begin{aligned} Z &= -(\partial V/\partial r) \\ &= -(1/a^2)[-(l+1)C_{lm} + lC'_{lm}]\mathrm{P}_{lm}(\cos\theta). \qquad (24.38) \end{aligned}$$

The different radial dependences of the primed and unprimed terms give opposite signs to the terms in Eq. (24.38), but the same signs in Eq. (24.37), allowing the ratio C/C' to be obtained from Z/X. This is the principle by which Gauss demonstrated that the main field is of internal origin, but we are considering here disturbance fields for which the internal components (the unprimed coefficients) must be explained by currents induced in the Earth by the external components. The radial variation in conductivity is obtained from the frequency variation of the ratio of internal to external fields.

In the approximation that the Earth is spherically symmetrical, the harmonic components of inducing and induced fields correspond, but there are complications. Field disturbances during magnetic storms are different on the sunlit and dark sides of the Earth. With sufficiently extensive data this can be accommodated by spherical harmonic analysis, but the induced field is subject to a phase delay, as in Eq. (24.33), and the Earth is rotating. The diurnal variation is particularly useful in this connection, because it has a very specific frequency, and there is a smaller but still well observed semi-diurnal variation, but these appear as waves propagating around the Earth, requiring a slightly different analysis. Greatest penetration is achieved with the longest period disturbances, which are essential to the estimation of lower mantle conductivity, but beyond a few months become confused with the secular variation. This limits the reliable information to the top part of the lower mantle, where a conductivity of about $1\,\text{S}\,\text{m}^{-1}$ is inferred. This is consistent with the average lower mantle value estimated above from the 1969 jerk.

We conclude that there is no dramatic radial variation in lower mantle conductivity, but temper this conclusion by noting the possibility of lateral heterogeneity, especially at the base of the mantle (layer D″). This is recognized in the conductivity profile in Fig. 24.5. A widely canvassed suggestion is that the deep mantle

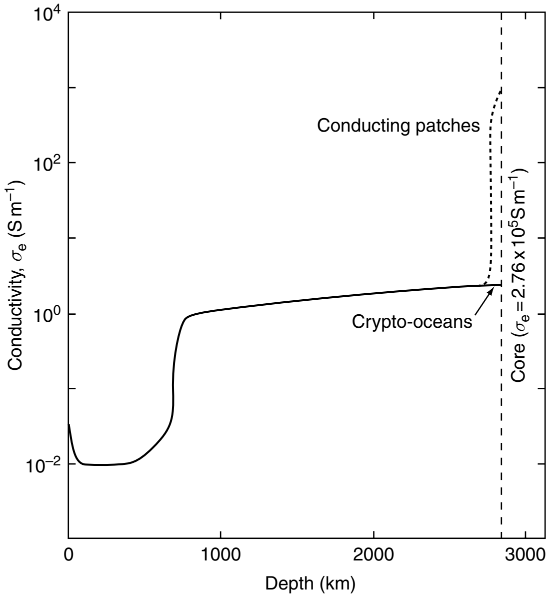


FIGURE 24.5 Electrical conductivity of the mantle.

under the Pacific Ocean is so much more conducting than the mantle elsewhere that it limits the penetration of the secular variation and the non-dipole field over a wide area centred on Hawaii (the Pacific 'dipole window'). The evidence is reviewed by Merrill *et al.* (1996, pp. 259–261), who conclude that the suggestion arises from a misinterpretation of paleomagnetic data and that the secular variation recorded by Hawaiian lavas is no different from that seen elsewhere. We note an argument by Buffett (1992) that a layer of metallic conductivity, with a thickness of a few hundred metres, at the base of the mantle would explain a phase lag in nutational motion. A complete layer of very high conductivity appears improbable and patches are more likely. They could have strong electromagnetic interaction with unseen fine-scale features of the field (or account for 'standing' features) but be small enough to have little influence on the global secular variation. This possibility is allowed in Fig. 24.5.

The mantle has a conductivity that is very low compared with that of metals, including the core. It must be classified as a semi-conductor. A feature of semi-conductors is an exponential dependence of conductivity on temperature. This was exploited by Dobson and Brodholt (2000) to infer that there is no major jump in temperature from the upper mantle to the lower mantle, unless the mineralogy has been misunderstood. This conclusion disallows the hypothesis of separate convective circulations in the upper and lower mantles because that would require a thermal boundary layer between them.

24.4 Electrical conductivity of the core

As far as we can tell, the Earth has always had a magnetic field, certainly for at least 3.5 billion years of its 4.5 billion year life. The geomagnetic dynamo is a robust feature and requires a core conductivity that is high enough to ensure that the dynamo is much more than marginally stable. The outer, fluid core must be a metallic conductor. There is no doubt or difficulty about this because available cosmochemical and geochemical evidence (Section 2.8) points to iron as the major core constituent. A necessary conclusion is that the thermal conductivity of the core is also characteristic of a metal, because thermal and electrical conductivities are linked by the Wiedemann–Franz law (Eq. (19.63)). There is, therefore, a continuous conductive heat loss by the core and this imposes a requirement on core energy sources that has provoked a vigorous debate over the need for a radioactive heat source. Stevenson (2003) pointed out that conductive heat loss is the most serious limitation on planetary dynamos. The chemical argument for potassium in the core is a subject of doubt (Section 2.8), and implicitly assumes that there is a strong physical argument for the need for it as a heat source. But that depends on the conductive heat loss and therefore on the estimate of conductivity, which is also seriously uncertain. Stacey and Loper (2007) argue that core conductivity is lower than has generally been supposed, perhaps low enough to avoid completely the need to assume radiogenic heat. The energy problem is a subject of Sections 21.4 and 22.7.

First, we consider the conductivity requirement of the dynamo. In Section 24.5 this is represented in terms of the dimensionless parameter, magnetic Reynolds number

$$R_m = Lv\mu_0\sigma_e, \tag{24.39}$$

where L is the scale size of motion at speed v, $\mu_0 = 4\pi \times 10^{-7}$ H m^{-1} is the permeability of free space and σ_e is conductivity. Dynamo action is favoured by increasing the size, speed of motion and conductivity; R_m is a product of all three. The argument leading to Eq. (24.45) (below) indicates that, for a sphere of radius L, a self-sustaining dynamo requires $R_m > \pi^2 \approx 10$. This is also the factor by which estimated electromagnetic relaxation times exceed the reorganization times for harmonic components of the field plotted in Fig. 24.4. We assume this limiting condition to apply to the smallest viable current loops in the core and that their size is indicated by the 160 km thick surface layer suggested by the spatial spectrum of the field in Section 24.2, that is, loops of cross-sectional radius 80 km. Putting $L = 80$ km in Eq. (24.39), with $v = 4 \times 10^{-4}$ m s^{-1}, as indicated by the speed of core fluid motion in the secular variation study by Bloxham *et al.* (1989), we have $\sigma_e \approx 2.5 \times 10^5$ S m^{-1}. Although rough and simplistic, this is a reasonable estimate of core conductivity and happens to be the mean of the values at the top and bottom of the outer core, as estimated below.

By the standard of metals, iron is a poor conductor. If the core were made of copper, with a conductivity more than 10 times higher, then core energy would be conductively dissipated too fast to permit dynamo action. In extrapolating the conductivity of iron to core conditions, it is helpful to understand the reason for the difference between copper and iron. An isolated copper atom has a single electron in its highest (4s) occupied state, with all inner shells filled, including a full complement of ten 3d electrons. When the atoms are combined in a metal, interactions between them spread the energy levels into bands, corresponding to the discrete shells of individual atoms. The lowest bands are filled to an energy level, known as the Fermi level, and the higher states are left vacant, except for thermal excitation of electrons between states within an energy range of order kT at the Fermi level. In metallic copper all the 3d states remain below the Fermi level and, as for the filled electron shells of insulators, cannot participate in conduction. The 4s band in copper is half-filled, with one electron per atom in the two available states. With the Fermi level in the middle of the 4s band, electrons close to it can readily change states, making copper a good conductor. The effective number of these mobile electrons decreases with pressure, as the band spread increases.

Iron has three fewer electrons per atom than copper and its 3d band is only partly filled, with the more widely spread, overlapping 4s band also partly filled. Both 4s and 3d electrons are available to contribute to conduction. Although the available 3d electrons are more numerous, with a high density of states at the Fermi level, they are more tightly bound and much less mobile. Conduction is dominated by the 4s electrons. It is limited by the scattering of electrons by phonons (lattice vibrations) or crystal irregularities, such as those caused by impurity atoms. The probability that an electron, accelerated by a field, will be scattered by any of these effects depends on the number of alternative states into which it can be scattered. For the 4s electrons in iron this number is greatly enhanced by the high density of 3d states. This is the reason for the high resistivity of iron, relative to copper. At low temperatures the effect of the 3d states is reduced by ferromagnetic alignment of their spins, invalidating any direct extrapolation from laboratory pressure–temperature conditions to the core. But the properties of iron at high temperature are well known, although extrapolation to core pressures requires some idea of the very different behaviour of the 4s and 3d states.

Experimental observations are also a subject of difficulty. Recent theories have been constrained by shock wave observations on iron and its alloys with Ni and Si up to pressures in the core range, but measurements on iron by Bi *et al.* (2002) have given much higher resistivities than earlier reports. Their samples were encapsulated in sapphire to avoid what they argued was a defect in earlier experiments, with the metal samples encapsulated in epoxy, which becomes conducting at a pressure of about 50 GPa. Shunting by epoxy biases low the apparent resistivity. This report prompted a re-examination of the theoretical arguments (Stacey and Loper, 2007).

The scattering of electrons by thermal vibrations can be regarded as a response to instantaneous deformation or irregularity of a crystal lattice. This increases with temperature but decreases with pressure, which reduces the amplitude of atomic vibration. Thermal disorder is essentially the same as that causing melting, and so we expect the effect of scattering on resistivity to be almost constant on the melting curve. Stacey and Anderson (2001) gave this argument a mathematical basis, concluding that it implies constant resistivity on the melting curve. However, this conclusion can apply only to electronically simple metals, such as copper, and not to iron, which has overlapping bands with different properties. In a simple metal with a single conduction band, the number of electrons available for electrical conduction, those within about *kT* of the Fermi level, is proportional to the density of states at that level. But the probability of scattering, being proportional to the number of available states into which an electron can be scattered, depends on the same number. Thus, to first order, conductivity does not depend on the density of states, provided it is high enough to give metallic conduction, and so is not materially affected by the increasing spread of the band by compression. In iron the 4s states, which dominate conduction, are spread much more than the 3d states and the effective number of conduction electrons decreases faster with pressure than the number of states into which they can be scattered. For this reason, pressure reduces the conductivity of iron more than it does for metals such as copper, in which all electrons at the Fermi level are of the same kind. A calculation by Bukowinsky and Knopoff (1977) indicated that at four-fold compression the entire 4s band would be above the Fermi level, in which case the conductivity would be very low, being due to 3d electrons only, but this is well beyond the terrestrial pressure range and is relevant only because it indicates a trend. Stacey and Loper (2007) made the simple assumption that the 4s states follow a relationship with the form of Eq. (19.17), but that the density of 3d states is much less affected by compression. On this basis they applied this equation as a multiplying factor to the resistivity of pure iron, as calculated by assuming that it is constant on the melting curve, and obtained 2.72 μΩ m at the top of the outer core and 3.75 μΩ m at the bottom.

The addition of impurity atoms to a solid metal introduces static irregularities to its lattice, causing an increment in resistivity that is independent of temperature and, for small concentrations, proportional to the impurity concentration. Thus, at atmospheric pressure a constant impurity contribution to resistivity becomes a decreasing proportion of the total as the resistivity increases with temperature. It is often supposed that the impurity effect can be neglected at high temperatures. However, although pressure reduces the thermal effect, it does not reduce the effect of impurity disorder, but, as experiments by Bridgman (1957) showed, increases it. With a sufficient impurity concentration the normal decrease in resistivity with pressure may be masked, and pressure and temperature both cause increased resistivity. Systematic measurements by Bridgman on a variety of iron alloys at pressures up to 10 GPa appear still to be the most relevant data that we have. They show that, for a variety of alloying elements, the break-even concentration, at which pressure has no effect on the resistivity of iron alloys at constant temperature, is about 14 atomic %. This is comparable to the concentration of light elements required to explain the core density (Section 2.8), so the effect of pressure on impurity resistivity in the core is probably slight, and we adopt the Stacey and Anderson (2001) value, 0.90 μΩ m.

Adding the impurity effect to the pure iron resistivity estimated above, we have a total resistivity of 3.62 μΩ m at the top of the core and 4.65 μΩ m at the bottom of the outer core. Corresponding conductivities are 2.76×10^5 $\mathrm{S\,m^{-1}}$ and 2.15×10^5 $\mathrm{S\,m^{-1}}$. For the inner core, with less impurity resistivity, the estimated conductivity is about 2.7×10^5 $\mathrm{S\,m^{-1}}$. As we note in Section 19.6, application of Eq. (19.63), with a small lattice contribution, gives thermal conductivities 28.3 W m^{-1} K^{-1} at the top of the core and 29.3 W m^{-1} K^{-1} at the bottom of the outer core. These values are used in assessing the core energy balance in Section 22.7, but it must be admitted that they are insecure. The electrical conductivity of the core is a prime target for further study.

24.5 The dynamo mechanism

Merrill *et al.* (1996, Chapters 8 and 9) give a comprehensive review of dynamo principles; this section selects salient features for comment. The underlying physics of the geodynamo is summarized in Fig. 24.6. The central idea is the frozen flux principle, illustrated by Fig. 24.6(a), which we have referred to in Sections 24.1 and 24.3. As we mention in connection with Eq. (24.39), and now consider more closely, if the combination of conductivity and the scale size and speed of motion (the magnetic Reynolds number, R_m, Eq. (24.39)) is high enough, then the field and the fluid move together. There is no inhibition to flow of fluid along field lines because that generates no electromagnetic forces; it is motion across field lines that is opposed by electromagnetic induction (Lenz's law). The motion would be completely prevented only in a superconductor, and in a normal conductor some diffusion of the field through the conductor occurs, generating electric currents that cause ohmic dissipation of energy. Dynamo action is possible only if the frozen flux principle prevails over field diffusion. As we demonstrate below, this requirement is satisfied if R_m exceeds about 10. An important point that is illustrated by Fig. 24.6(a) is that magnetic energy is created. The intensity of the field is represented by the closeness of the field lines, which is increased in the shear zone, where the total field is a vector superposition of the original field (upwards in the diagram) and a transverse field generated by the motion.

For dynamo action to be effective, the process of field regeneration, represented by Fig. 24.6(a), must be rapid enough to overcome the diffusion of the field out of the conductor. We can represent the two processes by time constants. For the regeneration process the time constant is

$$\tau_v = L/v, \tag{24.40}$$

where L is the scale of velocity shear with speed v, that is field lines can be moved over a distance L in the time τ_v if the relative speed is v. This must be less than the time constant τ_Ω for decay of the field by diffusion out of the conductor, that is, by ohmic dissipation,

$$\tau_\Omega = B/(-\mathrm{d}B/\mathrm{d}t), \tag{24.41}$$

where $(-\mathrm{d}B/\mathrm{d}t)$ is the rate of free decay of an unmaintained field, for example in a stationary

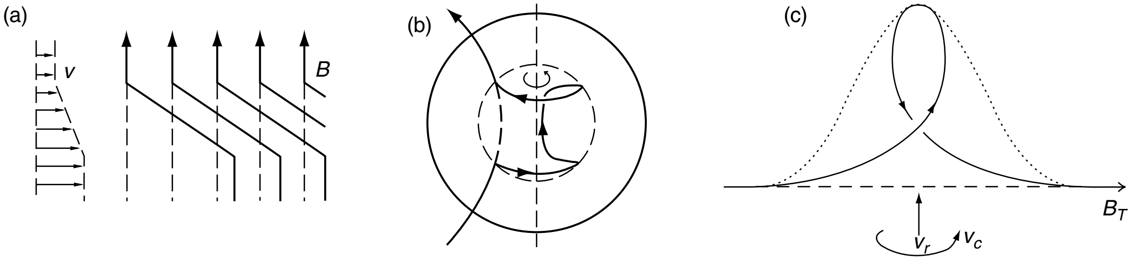


FIGURE 24.6 Basic physics of the dynamo mechanism, following the ideas of W. M. Elsasser. (a) A velocity shear perpendicular to a magnetic field in a fluid conductor deforms and intensifies the field. Field intensity is represented by the closeness of the field lines. (b) Differential rotation between the inner and outer parts of the outer core draws out the field lines of an initial poloidal field to produce an additional toroidal field in the shear zone. This is referred to as the ω-effect because it depends on differential rotation (at speed ω). (c) A toroidal field, B_T, is drawn out by a convective upwelling (at speed v_r) and, in the absence of rotation, would follow the dotted line. The Coriolis effect gives the fluid a helical motion, with rotational component v_c (anticlockwise in the northern hemisphere), deforming the field loop out of the plane of the diagram. This is the basis of the α-effect, generating a poloidal field from a toroidal one.

conductor. Thus, a necessary condition for dynamo action is

$$\tau_v < \tau_\Omega. \tag{24.42}$$

The free decay time-constant, τ_Ω, is an unambiguous quantity only for specific field patterns that maintain their forms as they decay in strength. We have a special interest in comparing the value of τ_Ω for the dipole field with the observed reorganization time, $\tau_1 \approx 1180$ years, given by Equation (24.26). Parkinson (1983, pp. 114–116) derived an expression for τ_Ω for the simplest dipole field pattern due to currents in a uniform sphere of radius a and conductivity σ_e,

$$\tau_{\Omega 1} = \mu_0 \sigma_e a^2/\pi^2 = (a/\pi)^2/\eta_m, \tag{24.43}$$

where

$$\eta_m = 1/\mu_0 \sigma_e \tag{24.44}$$

is referred to as magnetic diffusivity. Unlike R_m, this is a material property; its value in the core averages about 3.2 $m^2 s^{-1}$. To emphasize that this is a diffusion problem we point out that if η_m is replaced by thermal diffusivity, $\eta = \kappa/\rho C_P$ (Eq. (20.2)), then Eq. (24.43) gives the relaxation time for cooling of a sphere by thermal diffusion (Problem 19.2, Appendix J). The relationship between dissipation and magnetic energy depends on the current pattern and Eq. (24.43) applies to a pattern that remains constant as it decays. Substituting values of σ_e and a for the core in Eq. (24.43), we obtain $\tau_{\Omega 1} = 3.9 \times 10^{11}$ s $= 12\,200$ years. This is ten times the longest observed field reorganization time constant τ_1, as plotted in Fig. 24.4, satisfying Eq. (24.42).

Comparing τ_Ω, by Eq. (24.43), with τ_v, by Eq. (24.40) with $L = a$, the inequality in Eq. (24.42) becomes

$$\mu_0 \sigma_e a v > \pi^2. \tag{24.45}$$

The quantity on the left-hand side of this equation is the magnetic Reynold's number, R_m (Eq. (24.39)), with L identified as the radius of the sphere. As may be verified by substituting units or dimensions of its component parameters, it is a dimensionless number. The precise value that it must exceed to make a viable dynamo depends on geometrical details of the regenerative flow and boundary conditions and Equation (24.45) is only an approximate relationship. The competition between regeneration and diffusion is represented in a general way by the magnetic induction equation

$$\partial B/\partial t = \eta_m \nabla^2 \mathbf{B} + \nabla \times (\mathbf{v} \times \mathbf{B}). \tag{24.46}$$

The first term in this equation represents field diffusion and the second gives its interaction with material velocity that allows regeneration. Although electric currents are necessarily involved, and can be calculated from the B field, they do not appear in this equation, which simplifies dynamo calculations. Even so, Eq. (24.46) must be combined with expressions for the buoyancy and rotational forces driving **v** and the combination is not amenable to analytical solution. Numerical methods are required.

Two kinds of motion in the core are usually distinguished. The simpler to visualize is differential rotation, caused by the tendency of convecting fluid to conserve its angular momentum as it moves radially in response to buoyancy forces. This causes the inner part of the core to rotate with a higher angular velocity than the outer part and the shearing motion draws out the lines of a poloidal field to produce an additional toroidal field, as in Fig. 24.6(b). This mechanism is known as the omega (ω) effect. Although differential rotation within the core is central to dynamo theory, the supposition that it is steady is too simple and it may even reverse. Implications are considered further in Section 24.6. The convective radial motion is needed to regenerate the initial poloidal field from the toroidal one, by the second process, the alpha (α) effect, discussed by Roberts and Gubbins (1987) and Roberts (1987). An essential feature of the motion that drives the α-effect is its helicity, that is its spiral nature. Material drawn into an upwelling motion is deflected by the Earth's rotation (the Coriolis effect), causing cyclonic motion, as in the atmosphere (Fig. 24.6(c)). In the northern hemisphere rising material rotates anti-clockwise when viewed from above and the sign of this helicity is reversed for sinking fluid. Both signs are opposite in the southern hemisphere. It is important to recognize that the energy for dynamo action is derived from the convectively driven radial motion and that differential rotation cannot contribute to it.

A toroidal field line would be drawn out by the helical motion, as in Fig. 24.6(c), resulting in a field loop out of the plane of the diagram. Such a field loop is of poloidal form, but for generation of a large-scale poloidal field, such as the Earth's dipole field, non-zero average helicity is required. It is possible that this arises by a systematic latitude dependence of upwelling and downwelling convection, especially an asymmetry between the hemispheres, but non-linear effects arising from narrow upwelling with a broad-scale return flow would achieve the same result. A dynamo that relies on generation of a toroidal field from a poloidal one by the ω-effect and regeneration of the poloidal field by the α-effect is referred to as an α–ω dynamo. Dynamos that rely on the α-effect for both stages are also possible and are termed α^2 dynamos. It is possible for an α^2 dynamo, driven entirely by a repeated pattern of small-scale motions, as in turbulence, to generate a large-scale field.

Simple mechanical models have been useful in understanding how a self-exciting dynamo can operate. They include computer calculations of the behaviour of the coupled disc dynamo in Fig. 24.7, which was analysed by Rikitake (1966). Each of the discs rotates in an axial field produced by a coil carrying a current driven by the other disc. Rotation in the field generates an e.m.f. between the perimeter and the axis, providing the current that causes the axial field in the other disc. Self-excitation occurs when current generation overcomes ohmic dissipation by an adequate combination of size, speed and loop conductance. The model is symmetrical with respect to polarities of the currents and fields. If a particular combination of currents, and consequent fields, is self-exciting, then this is equally true when both currents (and fields) are reversed. Instability, including spontaneous current reversals, has been found with certain combinations of model parameters. Such time-varying behaviour depends on the loop inductances as well as the other model characteristics.

A practical realization of the principle of coupled disc dynamos was a laboratory model constructed by Lowes and Wilkinson (1963, 1968). The geometry of this device is illustrated in Fig. 24.8(a). Two metal cylinders were rotated at controlled speeds within cavities in a metal block, to which they were connected electrically by filling the gaps with mercury. No electrical or magnetic excitation was deliberately introduced, but ferromagnetic material was used for both the cylinders and the block, so that self-excitation could occur at attainable rotation speeds. The performance of the early version of the Lowes–Wilkinson model is illustrated in Fig. 24.8(b). The slight excitation at low speeds is attributed to residual magnetism in the metal, but the onset of self-excitation occurred quite sharply when a critical combination of rotation speeds was reached. A later, more sophisticated version of the model demonstrated instabilities, with spontaneous oscillations and reversals of the measured field. Use of ferromagnetic material in the Lowes–Wilkinson dynamo was necessary to suppress field diffusion sufficiently to allow dynamo action with a reasonable size and speed of the apparatus. The free space permeability μ_0 in the equation for magnetic diffusivity (Eq. (24.44)) is replaced by the much higher permeability of the magnetic material. However, ferromagnetism has no relevance

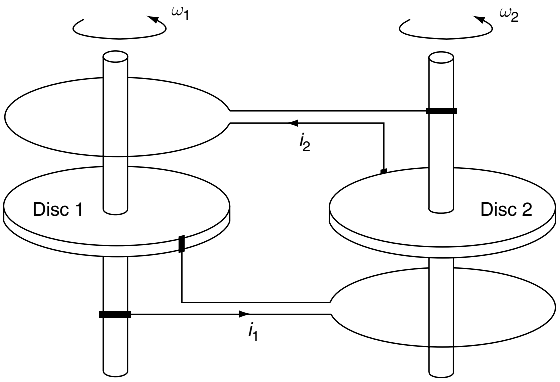


FIGURE 24.7 A simple mechanical analogue of the geomagnetic dynamo mechanism. This shows two interconnected disc dynamos. A single disc dynamo, which uses its current output to produce its own axial magnetic field, is self-excited if it has sufficient rotational speed. An interconnected pair, as shown here, may show more complicated effects, including spontaneous field reversals.

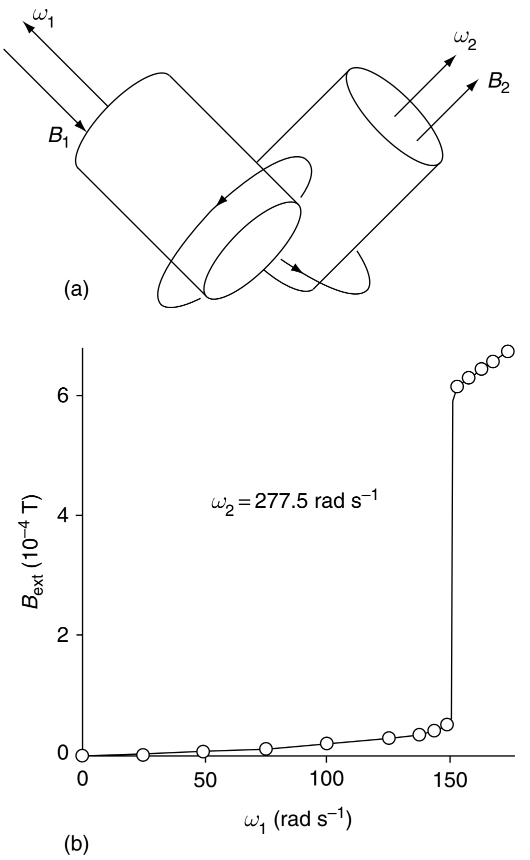


FIGURE 24.8 (a) Arrangement of rotating cylinders in the Lowes and Wilkinson (1963, 1968) laboratory model of geomagnetic dynamo action. The cylinders are enclosed in a metal block with which they are electrically continuous. (b) Externally observed field as a function of the speed of cylinder 1, with cylinder 2 maintained at a fixed speed.

to the Earth's core where the permeability is insignificantly different from μ_0. Progress on more recent laboratory experiments, with non-magnetic media, mostly liquid sodium, has been discussed in a special journal issue (Rädler and Cēbers, 2002). A significance of such experiments is that, by using a conducting liquid of low viscosity, the role of turbulence can be examined in a way that is not possible in numerical dynamos.

The theoretical approach to self-exciting dynamo action in a fluid conducting sphere took an important step forwards with the model by Bullard and Gellman (1954). They applied the principle of α–ω dynamo action by representing poloidal and toroidal fields, with their associated motion, as spherical harmonics feeding one another. This work was influential in focussing attention on several features that became central to discussions of the dynamo: the necessity for a toroidal field, which is not observable, poloidal–toroidal field interaction and differential rotation within the core. The success of their approach was limited by lack of convergence of their harmonic series. This fact is interesting in view of the evidence that the spatial spectrum may be white and that the various harmonics cannot occur independently, with energy fed down the harmonic chain. The Bullard and Gellman approach pioneered kinematic dynamo modelling, that is the development of models in which the motion is specified and the electromagnetic consequences are calculated. The disc dynamo and the Lowes–Wilkinson laboratory model formally come into this category, although with less Earth-like geometry.

Fully dynamic numerical modelling, with fluid motion impelled by buoyancy and self-adjusted by the mutual control of the motion and the field, became possible with very large and fast computers. There are different ideas about assumptions and approximations (Glatzmaier and Roberts, 1995a,b; Kuang and Bloxham, 1997; Takahashi *et al.* 2005) that yield internal motions and field patterns that may be quite different from one another in detail, but they generally give quite Earth-like behaviour, with dominant dipole fields. Kono and Roberts (2002) reviewed the various approaches. Dominance of the dipole field is evidently a feature of the terrestrial situation but, as modelling reported by Busse (2002) shows, different physical conditions may lead to quadrupole dominated dynamos. This appears to be the case in Uranus and Neptune. In using the models to judge what is important in the physical assumptions, by comparing them with observations of the Earth's field, we are limited by the fact that the fine details are obscured by our distance from the source. Also our direct observations extend over a very short time, compared with the time scales of geomagnetic effects, and we turn to paleomagnetic evidence, discussed in Chapter 25.

First, we note that the fluid viscosity of the outer core is almost certainly quite low, within a factor 10 or so of the viscosity of water (Poirier, 1988; Dobson, 2002). This means that viscous forces are very weak compared with the Coriolis force of rotation and in this situation convective motion tends to break up into quasi-independent columns (Taylor columns) parallel to the rotation axis, with little viscous interaction between them. The 'tangent cylinder', a cylindrical surface parallel to the axis and just enclosing the inner core, assumes particular significance. Although not a physical boundary, in many models it effectively isolates core motion in polar regions (within the tangent cylinder) from the rest of the outer core. But molecular viscosity may not really be relevant to the motion of a magnetically controlled fluid, in which eddy viscosity has a much higher value and is anisotropic, being controlled by the field. The possibility that the tangent cylinder has an observable surface expression is examined by Olson and Aurnou (1999). What evidence do we have for Taylor columns in the core? Correlated 'flux bundles' in opposite hemispheres (Bloxham, 2002) are probably an indication of their existence. A related question is the rotation of the inner core. Differential rotation within the outer core, with the inner core rotating slightly faster than the mantle, is a feature of several theories and was drawn to attention by the model of Glatzmaier and Roberts (1996). Reports that it is observed seismologically are discussed in Section 24.6, but the Kuang and Bloxham (1997) model gives inner core rotation that may be either faster or slower at different times and gravitational interaction between the inner core and the mantle complicates the picture (Section 24.6).

Reversals of the dipole are a common feature of the models and the mechanism is of particular interest. In the model of Takahashi *et al.* (2005) reversals appear to be initiated by flux patches that start at low latitudes and migrate to higher latitudes. The latitude migration of sunspots with their intense fields, during the 22-year sunspot cycle of solar field reversals, appears similar although it occurs in the opposite direction. But terrestrial field reversals are very irregular and have a long-term frequency variation that is suggestive of mantle control (Jones, 1977; McFadden and Merrill, 1984 – see Section 25.4). It is feasible that flux patches of the kind seen in the Takahashi *et al.* (2005) model would be triggered by thermal structure in the D″ layer at the base of the mantle. Variations in this structure on the 10^8-year time scale of mantle convection appear to offer a plausible explanation for the very variable rate of reversals (Section 25.4).

The intervals between reversals, typically several hundred thousand years, are much longer than the duration of the reversal process, which may be 5000 years or even less. The inner core may have a stabilizing effect, inhibiting reversals (Hollerbach and Jones, 1995), because the magnetic flux in the inner core can change only by diffusion and its relaxation time, about 1600 years, is longer than the reorganization times plotted in Fig. 24.4.

Although many details of the dynamo remain obscure, or subject to different theories, the general principles are not in doubt. Any sufficiently large-scale and complicated motion of a fluid conductor will generate a field. Like any other state, the zero field state is unstable and cannot occur if conditions favour field generation. It is useful to anticipate two results of paleomagnetism (Chapter 25) by appealing to Curie's principle of symmetry, according to which no effect can have lower symmetry than the combination of its causes. If we assume that, apart from the effect of rotation, the core and the compositional and thermal gradients within it are spherically symmetrical, then the rotational axis must be a symmetry axis in the operation of the dynamo, and the only one. Curie's principle requires that, averaged over a sufficient multiple of its relaxation time, the magnetic axis must coincide with the rotational axis. This is recognized as the axial dipole hypothesis in paleomagnetism (Section 25.3). Moreover, nothing in the dynamo distinguishes the two opposite directions of the axis. As we now know from paleomagnetism (Section 25.4) the field may have either polarity with equal probability. Curie's principle is discussed in the context of inner core anisotropy in Section 17.9, where it is pointed out that in geophysical or geological contexts of this kind it must be applied in a statistical sense. An analogy is a turbulent stream, in which individual drops of water have highly variable motions, but with sufficient averaging, either over all the drops or over time for one drop, are seen to move steadily downstream. Provided this statistical approach is adopted, the symmetry principle is a powerful constraint on our physical theories.

24.6 The westward drift and inner core rotation

The historical record of the secular variation in western Europe extends back to about 1600. The 400 years of data from London (Fig. 25.3) appear as a rotation of the field direction about the inclined dipole field. Bullard *et al.* (1950) noted that this is consistent with a westward drift of the non-dipole features (Fig. 24.1) past any point of observation. It gives the impression that the westward drift is a permanent feature of the secular variation and it was interpreted by Bullard as a bodily motion of the outer part of the core relative to the mantle. The spectrum of the length of day observations (Sections 7.4 and 7.5) indicates a relaxation time for establishment of rotational equilibrium between the core and mantle that is a few decades at most and this is incompatible with prolonged differential rotation. For this reason, Bullard postulated that the mantle was somehow coupled to the deeper seated dipole field and that the westward drift was a direct indication of differential rotation within the core. There is a fundamental difficulty with this interpretation. Electromagnetic coupling of the core and mantle is due to the field at the outer boundary of the core and, to the extent that the frozen flux principle applies, the entire field moves with the outermost part of the core. The mantle cannot be coupled independently to a deep feature of the field, which can enter the mantle only through the surface layer of the core. An obvious, but contrived, explanation is that we do not see all of the field at core level and that the westward drifting long wavelengths are compensated by unseen finer scale features with a net eastward drift. Subsequent extension of the record by archeomagnetic measurements to span 2000 years, as in Fig. 25.3, shows that the simple cyclic pattern is characteristic only of the last few hundred years and not the earlier period. But a systematic drift, even for a few hundred years, has appeared paradoxical.

Figure 25.3 represents the field in a small local area, whereas the drift of the features in Fig. 24.1 is global. Moreover, the non-dipole field is not only drifting but its pattern is changing, so that Fig. 25.3 is not a conclusive argument against the significance of the drift. If the core–mantle coupling coefficient, $K_R = 3.2 \times 10^{28}$ N m s (Eq. (7.33)), is correctly estimated from the spectrum of length of day observations, with a relaxation time of about 8 years, and the westward drift is $\Delta\omega = 0.18^\circ/\text{year} = 1.0 \times 10^{-10}$ rad s^{-1}, the energy dissipation by a bodily drift at this rate would be $K_R \Delta\omega^2 = 3 \times 10^8$ W. This is small enough to present no difficulty, energy-wise, but it would preclude differential rotation lasting for hundreds of years. The essential difficulty with the drift is in devising a mechanism to maintain it against the electromagnetic coupling that would establish equilibrium between the rotations of the mantle and the top of the core. A gravitational interaction, considered below, appears to be the only possibility.

Independent evidence for differential rotation is faster rotation (super-rotation) of the inner core. This was drawn to attention by the model of Glatzmaier and Roberts (1995a,b) and seismological evidence was first presented by Song and Richards (1996) (see also Section 17.9). The basis of their observation is that the inner core is not completely homogeneous and is anisotropic, with an anisotropy axis close to but not coinciding with the rotational axis, so that any rotation of the inner core, relative to the mantle, gives a time-dependence to the travel times of seismic waves with paths that enter the core. The effect is small and its interpretation requires detailed information on the structure of the inner core. The rate of relative rotation was estimated by Song and Richards (1996) to be about 1°/year, comparable to the value for the Glatzmaier and Roberts (1996a,b) model, although that varied with time, as did the relative rotation for the Kuang and Bloxham (1997) model. Further studies yielded both widely differing rotation rates and refutations (Su *et al.*, 1996; Creager, 1997; Souriau *et al.*, 1997; Laske and Masters, 2003). A later, more secure estimate is 0.3 to 0.5°/year (Zhang *et al.*, 2005).

Meanwhile, Buffett (1996) suggested that inner core super-rotation might not be expected, because of gravitational coupling to heterogeneities in the lower mantle. The idea is that the gravity field of the heterogeneities would cause the development of topography on the inner core boundary and that this would be 'locked' to the gravity field that caused it. While it is plausible that the gravitational torques would be strong enough if such topography exists, its development requires special conditions. If the inner core, starting from nothing and growing at 0.3 mm/year (1200 km in 4 billion years), is rotating at (say) 0.5°/year, it would have grown by 5 cm in radius in the 180 years required for a 90° rotation, and this is far too little to represent significant topography. The rotation would smear out any tendency to uneven growth at locations fixed relative to the mantle. The required topography could develop only if the inner core is soft enough to deform in response to the gravitational forces, which would then resist, but not stop, the rotation, because the deformation would continuously readjust, a process analogous to tidal friction (Section 8.3). The possibility of a drag by gravitational interaction with the mantle threw a lifeline to Bullard's notion of coupling of the mantle to the deep core, allowing rotation of the top of the core, relative to the mantle, and therefore a westward drift, to persist indefinitely. It appears that the top of the core has been rotating more slowly than the mantle for several hundred years, but that this is a consequence of the coupling of the mantle to the inner core and is probably not a permanent feature. The inner core is coupled to irregular motion in the inner part of the outer core, so that the westward drift may come and go according to variations in the convective pattern very deep in the core (see also Section 17.9).

Inner core rheology is important also to the interpretation of its anisotropy, which Yoshida *et al.* (1996) explained by preferential deposition of solid in equatorial regions and anelastic relaxation of the inner core towards equilibrium ellipticity. The anisotropy is due to crystalline alignment caused by the continuing slow deformation. This means that the inner core is continuously relaxing from a state of excess ellipticity. If the excess can be reliably observed then the relaxation time and the implied viscosity can be estimated. Such an estimate would impose a tight constraint on the theory of gravitational locking (Buffett, 1997).

We are led to a tentative conclusion that the relative motion of the inner core and mantle is variable but that 'super-rotation' predominates. Nevertheless, gravitational coupling may exert a drag on the inner core rotation and provide a torque on the mantle sufficient to allow a persistent westward drift in spite of the opposite electromagnetic torque imposed by the top of the core. If this is so, then it raises doubt about the validity of the core–mantle coupling coefficient (Eq. (7.33)), which assumes only electromagnetic coupling of the core and mantle.

24.7 Dynamo energy and the toroidal field

In Section 24.2 we suggest that the rms strength of the poloidal field in the core is about 11.8×10^5 nT. This corresponds to a magnetic energy density

$$(E/V)_{\text{Poloidal}} = B_P^2/2\mu_0 = 0.55\ \text{J m}^{-3}. \qquad (24.47)$$

For any plausible speed of core motion, this is much greater than the kinetic energy,

$$(E/V)_{\text{Kinetic}} = \rho v^2/2 = 8 \times 10^{-4}\,\text{J m}^{-3}, \qquad (24.48)$$

assuming $\rho = 10^4\ \text{kg m}^{-3}$ and $v = 4 \times 10^{-4}\ \text{m s}^{-1}$. If the energy of the toroidal field is added, the difference is even greater. These numbers demonstrate that mechanical inertia has little influence on the dynamo mechanism and that the inertia of the system is due to the magnetic field (a magnetic field imparts inertia to the electric current producing it). The convective forces driving core motion work directly against the magnetic forces and kinetic energy is irrelevant. Also, the molecular viscosity is very low, allowing us to conclude that the convective energy calculated in Section 22.7 is converted almost entirely to magnetic energy, which is dissipated by ohmic heating, with no other limitation or inefficiency. Thus, by subtracting the ohmic dissipation by the poloidal field from the estimated total convective power, we have an estimate of the dissipation by the toroidal field. From this we can estimate the ratio of toroidal to poloidal field strengths.

Consider first a current of uniform density i circulating about the magnetic axis in a sphere of radius a. Its dipole moment is (Problem 24.2)

$$m = \frac{\pi^2}{8} a^4 i. \tag{24.49}$$

If we take a as the core radius with m given by Eq. (24.2), then we require $i = 4.3 \times 10^{-4}\,\mathrm{A\,m^{-2}}$. The ohmic dissipation is i^2/σ_e per unit volume, that is the total dipole field dissipation is

$$\begin{aligned}\left(-\frac{\mathrm{d}\varepsilon}{\mathrm{d}t}\right)_{\text{Dipole}} &= \frac{i^2}{\sigma_\mathrm{e}}\frac{4}{3}\pi a^3 = \frac{256}{3\pi^3}\frac{m^2}{\sigma_\mathrm{e} a^5} \\ &= 1.3 \times 10^8\,\mathrm{W}. \end{aligned} \tag{24.50}$$

We can use this as a guide to the total dissipation.

The spectrum of the field, as represented by Fig. 24.2, leads to the conclusion that the rms field strength of all harmonic components of degree l, taken together, is independent of l up to a limit which we assess as $l_{\max} = 136$. If the higher degree components were identified with independent current loops of corresponding sizes, then the required currents would be inversely proportional to loop size. But this is not relevant. The current loops are linked in a network, with a continuity of magnetic flux passing between them. The field at any point is due not just to an adjacent loop, but to an inseparable combination of them all, with flux paths having some resemblance to the domains in ferromagnetic material (Section 25.2). The result is that a white field spectrum is produced by a white current spectrum. So, we assume the rms current density, and therefore ohmic dissipation, by harmonics of degree l to be independent of l for $l \geq 2$.

The dipole field is stronger than the trend of the other harmonics in Fig. 24.2 and Eq. (24.24) by a factor of about 6 in spectral energy (or $\sqrt{6}$ in rms field strength). If the rms current density is the same for all higher harmonics, they each dissipate $2.5 \times 10^7\,\mathrm{W}$. If there is a total of n harmonics, the total dissipation is

$$\begin{aligned} -\left(\frac{\mathrm{d}\varepsilon}{\mathrm{d}t}\right)_{\text{Poloidal}} &= -\left(\frac{\mathrm{d}\varepsilon}{\mathrm{d}t}\right)_{\text{Dipole}}\left(1 + \frac{n-1}{6}\right) \\ &= 2.6 \times 10^9\,\mathrm{W}, \end{aligned} \tag{24.51}$$

where $n = 136$ is adopted from Section 24.2. If the toroidal field is about ten times as strong as the observed poloidal field, making it about 5×10^6 nT (5 mT), it is comparable to the estimate of the radial field at the inner core boundary, 7.2×10^6 nT, by Mathews *et al.* (2002) from the electromagnetic coupling required to explain observations of nutations. Oscillatory motions of the Earth driven by gravitational interactions with other bodies, termed nutations, have amplitudes and phases that are influenced by the internal motions that they cause. As an example, the nutation accompanying precession is referred to in Section 7.2 and indicated in Fig. 7.6. The flux linkage between the inner and outer cores must be dominated by the toroidal field, which has, at that level, strong radial components penetrating the inner core. This is to be expected because the inner core is a slightly better conductor than the outer core and would carry the necessary currents. Thus, the nutation calculations indicate a toroidal field 10 times as strong as the poloidal field, implying ohmic dissipation 100 times as great, that is 3×10^{11} W. This is the dynamo power requirement assumed in Section 21.4.

It remains to re-examine the role of precession as a core energy source. In Section 22.7 we mention the estimate of the present energy dissipation in the core by precessional torques, 6×10^9 W. This is lost in the uncertainties in our calculation of core energy, but that may have been less true in the remote past, when the Moon was closer and the Earth was rotating faster and so was more elliptical. Our reason for re-opening the question here is to assess the plausibility of a significant energy contribution to the dynamo early in the life of the Earth.

First, we consider the present situation, following some of the arguments of Stacey (1973). The rate of change in the orientation of the Earth's rotational axis on its precessional path is $2\pi \sin\theta/\tau$, where $\theta = 23.45^\circ$ is the angle between the rotational axis and the ecliptic pole, around which it rotates in a period $\tau = 25\,800$ years. The core axis must keep up with this, but only ¾ of the necessary torque comes from the lunar and solar gravitational interactions, because the core is denser and therefore less elliptical than the Earth as a whole. So, core–mantle interaction must provide a torque sufficient to change the orientation of the angular momentum axis of the core at a rate

$$0.25 \times 2\pi \sin\theta/\tau = 7.7 \times 10^{-13} \text{ rad s}^{-1}. \quad (24.52)$$

If the core rotation axis is left behind by an angle α then its rotation is misaligned by this angle relative to the symmetry axis of the cavity. The mantle exerts a precessional torque on the core, in the manner of Toomre's marble analogy in Section 7.5, and the angular rate of its precessional motion is $fe\omega$, where $e = 2.45 \times 10^{-3}$ is the ellipticity of the cavity, ω is the rotation rate and $f \approx 0.7$ is a factor to allow for the elastic distortion of the cavity. We assume this to have the same value as the factor f in Eq. (7.24) that accounts for the lengthening of the Chandler wobble period. This motion causes the core axis to move to a wider angle of precession, as in Fig. 7.6, with α self adjusted so that the torque exerted by the mantle makes up for the deficiency in the lunisolar torques. The two axes precess together, separated by the small angle α. Thus

$$fe\omega\alpha = 7.7 \times 10^{-13} \,\text{rad s}^{-1}, \quad (24.53)$$

requiring

$$\alpha = 6.1 \times 10^{-6} \,\text{rad}. \quad (24.54)$$

Equation (24.54) gives the misalignment of the core and mantle axes with the assumption that the coupling is entirely inertial. The dissipation implied by this angle if electromagnetic coupling is superimposed is

$$-\mathrm{d}E/\mathrm{d}t = K_\mathrm{R}(\omega\alpha)^2 = 6.4 \times 10^9 \text{ W}, \quad (24.55)$$

where K_R is the coupling coefficient given by Eq. (7.33). The addition of dissipative coupling introduces the lag angle δ in Fig. 7.6. It is essential to the argument that the coupling is mainly inertial because, if it were entirely dissipative, the angular separation of axes would be much larger and cause dissipation of about 8 terawatts.

Now we are in a position to assess the effect of faster early precession. By Eq. (7.11), the angular rate of precession, Ω, varies as $(C-A)/R^3\omega$, where R is the distance to the Moon and $(C-A) \propto \omega^2$. Thus $\Omega \propto \omega/R^3$ and the product $(fe\omega\alpha)$ in Eq. (24.53) is increased by this factor. But, like $(C-A)$, e is proportional to ω^2, so, with f assumed to be constant, $\omega^3\alpha \propto \omega/R^3$, or

$$\omega\alpha \propto 1/\omega R^3. \quad (24.56)$$

Then, by Eq. (24.55), the dissipation varies as $1/\omega^2 R^6$. We do not need to consider the Moon to be closer than 30 Earth radii (half of the present distance). With conservation of angular momentum in the Earth–Moon system, this means a 10 hour rotation period, that is ω greater than at present by a factor 2.4. With these numbers, the dissipation in Eq. (24.55) would be increased by a factor 11 to about 7×10^{10} W. Bearing in mind that this is mechanical, not thermal, energy, it is big enough to be interesting. Precession could have provided a significant contribution to dynamo power early in the life of the Earth, when the inner core was small, or non-existent, and compositional convection ineffective, although some ohmic dissipation would have occurred in the mantle.

The westward drift has sometimes been attributed to the small angular difference between the core and mantle rotational axes, but the value of α in Eq. (24.54) is quite inadequate. The equilibrium difference between the rotation rates is $\Delta\omega/\omega = 1 - \cos\alpha \approx \alpha^2/2 = 2 \times 10^{-11}$, giving $\Delta\omega = 1.5 \times 10^{-15}$ rad s$^{-1} = 2.7 \times 10^{-6}$ degrees/year. A westward drift of 0.2°/year would require $\alpha = 0.1^\circ$ and then Eq. (24.55) would give an energy dissipation of 5×10^{14} W.

24.8 Magnetic fields of other planets

Magnetic fields have been sought on the Moon and planets (Table 24.2). The giant planets all have active dynamos, made possible by the conversion of their light gaseous components to metallic forms by the very high internal pressures. Table 24.2 lists estimated dipole moments and corresponding mean surface field strengths, but in several cases, especially Uranus and Neptune, the field is far from dipolar and these numbers must be taken as no more than an indication. Among the terrestrial planets, and including the Moon, the Earth is exceptional. Their fields are all much weaker and, with the probable exception of Mercury, require no appeal to currently active dynamos, but are apparently due to magnetizations of their crusts by former fields. In the case of Venus, the crust is too hot to retain magnetic remanence, and it probably has no field at all. Mars is an oxidized planet with a high iron oxide content of its crust. It has a strong crustal magnetization, giving an irregular field for which the estimated dipole moment is irrelevant as there is evidently no dynamo. The magnetic pattern is apparent at satellite altitudes and includes magnetic stripes (Connerney *et al.*, 1999), apparently similar to those on the ocean floors. This is intriguing evidence not just of an early dynamo, now extinguished, but also of magnetic reversals and mantle convection of plate tectonic form. All of the terrestrial planets and the Moon have cores that are still at least partly liquid and they would almost certainly have had dynamos early in their lives, a conclusion reached by Stevenson (2003) in a review of planetary magnetism.

The magnetic field of Mercury is of particular interest and merits more detailed attention. Its strength is in the middle of the range of the planetary fields, being much stronger than the fields of the planets that clearly lack dynamos but very weak compared with the fields of the Earth and giant planets that certainly have dynamos. The evidence that the large core of Mercury is still at least partly liquid (Margot *et al.*, 2007) allows the possibility of an active dynamo, but an explanation in terms of crustal magnetization has not been totally abandoned (Aharonson *et al.*, 2004). The hypothesis that the field is a fossil relic of an early dynamo faces several difficulties. First, there is the problem that a uniform crust magnetized by an internal dipole would itself have no dipole moment. The magnetization of the equatorial crust would be opposite to that in the polar regions and precisely cancel it. But, on a planet so close to the Sun, the crust would not be magnetically homogeneous. The equatorial crust would be hot, possibly sufficiently so to demagnetize it thermally over geological time, leaving the polar regions to give a net dipole moment. Even so, the field strength is high enough to demand abnormally magnetic rock, or else a remarkably strong former dynamo. Mars has a strongly magnetic crust but its field is very weak compared with that of Mercury, and Mercury is evidently in a reduced state with little iron oxide in its mantle or crust. Moreover, we could hardly expect the field to remain as a steady dipole, systematically magnetizing the whole crust over a very long time. It would have drifted and probably reversed, giving scattered magnetization, as seen on Mars. Thus, crustal magnetization cannot be regarded as a serious candidate source for the field of Mercury.

We now consider the energy implications of a currently active dynamo in Mercury. Being a small planet, with gravity correspondingly

Table 24.2 Magnetic fields of planets. Data from Ness (1994) with a later revision for Mars

Planet	Dipole moment (Am^2)	Dynamo?	Surface field (nT)
Mercury	5×10^{19}	Probably	475
Venus	$<4 \times 10^{18}$	No	<2.5
Earth	8×10^{22}	Yes	41 455
Moon	$<1 \times 10^{16}$	No	<0.26
Mars	$\sim 1 \times 10^{18}$	No	3.5
Jupiter	1.6×10^{27}	Yes	650 000
Saturn	4.7×10^{25}	Yes	32 850
Uranus	3.8×10^{24}	Yes	32 170
Neptune	2.0×10^{24}	Yes	18 560

weaker than in the Earth, the thermodynamic efficiency of thermal convection is lower, as is also the energy that might be available from compositional separation. There is, however, a compensating advantage: it means a smaller adiabatic temperature gradient and less conductive heat loss. We can examine the balance of these effects by adjusting the core analysis in Sections 21.4, 22.6 and 22.7 to parameters appropriate to the core of Mercury. To focus on a plausible situation we adopt a Mercury core model investigated by Stanley *et al.* (2005), with an inner core that has grown to 0.8 of the core radius. This has the advantage that, if compositional convection still operates, it maximizes the ratio of compositional to thermal energies, allowing for dynamo action with a minimum cooling rate (as pointed out for the Earth's core in Section 22.7). We may wonder whether, at such an advanced stage of inner core development (55% of the core mass solidified), the outer core can still accommodate more rejected solute. But we find that, if it does so, then what Stanley *et al.* (2005) refer to as a thin shell dynamo is energetically viable.

Using the model of Mercury in Table 1.2, with the core equation of state parameters in Table 5 of Stacey and Davis (2004), the core properties at the core–mantle boundary pressure of Mercury (6.8 GPa) are: $\gamma = 1.74$, $K_S = 158\,\mathrm{GPa}$, $g = 4.09\,\mathrm{m\,s^{-2}}$, $\rho = 6890\,\mathrm{kg\,m^{-3}}$ and $T = 2200\,\mathrm{K}$ (by adiabatic extrapolation from the melting point at the inner core boundary, 383 km deeper). Applying these values to Eq. (19.55), we estimate the adiabatic temperature variation at the top of the core to be $\mathrm{d}T/\mathrm{d}z = 6.83 \times 10^{-4}\,\mathrm{K\,m^{-1}}$. We need to estimate also the thermal conductivity, by applying the Wiedemann–Franz law (Eq. (19.63)), as in the case of the Earth's core (Section 24.4). At the lower pressure of Mercury the modification of electron band structure by pressure is much less than in the case of the Earth and, for pure iron, the resistivity is nearer to the zero pressure liquid iron value ($1.35\,\mu\Omega\,\mathrm{m}$). With a small pressure effect and adding impurity resistivity, as for the Earth's core, a calculation similar to that in Section 24.4 gives $2.4\,\mu\Omega\,\mathrm{m}$ for the Mercury core, making it a slightly better electrical conductor than the Earth's core. But, in calculating the thermal conductivity by Eq. (19.63), the lower temperature in Mercury more than compensates for this difference and, with a small lattice contribution we estimate $\kappa = 25\,\mathrm{W\,m^{-1}\,K^{-1}}$. With this value we calculate the conducted heat flux at the top of the core:

$$\mathrm{d}Q/\mathrm{d}t = 4\pi r^2 \kappa \mathrm{d}T/\mathrm{d}z = 7.85 \times 10^{11}\,\mathrm{W}. \quad (24.57)$$

As in the Earth, the conducted heat is a base load on core energy sources. It amounts to $1.1 \times 10^{29}\,\mathrm{J}$ over 4.5×10^9 years. We can compare this number with the heat provided by core cooling and latent heat of inner core solidification, ignoring for the moment the possibilities of compositional convection and radiogenic heat. The calculation is essentially the same as that for the Earth in Section 21.3, replacing the numbers with those appropriate for Mercury. The specific heat of Mercury's core is somewhat smaller than for the Earth's core because, at the lower temperature, there is a smaller electron contribution, although the density of electron states is higher at the lower pressure (Eq. (19.17)). Also, in the smaller planet, less gravitational energy is released by contraction. Allowing for these effects, referred to the core–mantle boundary temperature as the core cools adiabatically, the effective heat capacity, that is the heat lost per degree cooling of the boundary, is $2.3 \times 10^{26}\,\mathrm{J\,K^{-1}}$. By an analysis similar to that for the Earth's core, leading to Eq. (21.19), the core–mantle boundary temperature drop accompanying development of the inner core to 0.8 core radii is 185 K, giving a heat release of $4.3 \times 10^{28}\,\mathrm{J}$. To this we add the latent heat of solidification by Eq. (21.11). The inner core mass, 1.24×10^{23} kg, is assumed to have the same mean atomic weight as the Earth's inner core ($m = 50.16$), so that $n = 2.47 \times 10^{24}$. The temperature at the solidifying interface varies from 3020 K to 2430 K as the inner core grows and the weighted average is close to the simple average, 2700 K. Using the Grüneisen parameter (Eq. (19.1)) to calculate the αK_T term, and Eq. (19.54) for the volume change accompanying solidification, we estimate the latent heat release to be

$$L = T_\mathrm{M} n R \ln 2(1 + 2\gamma\alpha T_\mathrm{M}) = 5.1 \times 10^{28}\,\mathrm{J}, \quad (24.58)$$

making the total heat release by cooling and freezing 9.4×10^{28} J.

We see that, with no allowance for compositional separation or radiogenic heat, the heat released during inner core formation does not fall far short of the conducted heat in the life of the planet. It is a much larger fraction of the conducted heat than in the case of the Earth's core but, of course, that is a consequence of the fact that the Earth's inner core is only 5% of the total core mass, whereas for Mercury we are considering 55%. Thus, a dynamo driven thermally, with an inner core age significantly less than that of the planet, is plausible. Although the average thermodynamic efficiency is modest (7%), so is the energy demand by the weak dynamo. Compositional separation may also have a role, although we suggest in Section 1.14 that, since Mercury is a reduced planet, we would not expect a large oxygen content in its core. But a modest compositional energy would permit consideration of an inner core no younger than the planet and still allow a weak dynamo. Planetary dynamos are generally believed to require rotation and the rotation of Mercury is very slow, but we may wonder whether, in spite of conventional disbelief, the solar tide has a role. The orbit of Mercury is very elliptical ($e = 0.2056$) and gives a strong 88 day radial tide that takes effect as a modulation of the 176 day angular tide. Although we still know rather little about Mercury, we conclude that its magnetic field is most plausibly attributed to a core dynamo.

The magnetizations of meteorites, and by inference of the asteroids, are discussed in Section 1.11. They are attributed to a former strong solar magnetic field. Mention should be made also of evidence for fields on Jupiter's satellite Ganymede and on the asteroids Gaspra and Ida, but these remain to be interpreted. Magnetic fields are ubiquitous, not just in the Solar System but in the Universe generally. Magnetohydrodynamic action occurs spontaneously in any conductor that is large enough and has sufficient vorticity in its internal motion.

25

Rock magnetism and paleomagnetism

25.1 Preamble

Paleomagnetism (ancient magnetism) is the study of the pre-history of the Earth's magnetic field, extending the record from the 400 years of direct observation to almost 4 billion years. It relies on the fact that many rocks retain indefinitely the remanent magnetism that is induced in them during their formation, and so record the direction and, with less certainty, the strength of the inducing field. That rocks have magnetic properties has been known since ancient times. By the late nineteenth century measurement methods had become sensitive enough to observe remanence in a wide range of rock types. However, they were not used to infer the history of the Earth's field. Even the discovery by P. David and B. Brunhes early in the twentieth century of rocks with magnetic remanence opposite to the Earth's field did not attract attention commensurate with its significance for another 50 years. The study of rock magnetism began to develop rapidly in the 1950s, when the revolutionary discoveries of paleomagnetism were appearing and the need for a fundamental understanding had become urgent. The two principal landmark publications in the early development of our understanding are Nagata (1953) and Néel (1955). Dunlop and Özdemir (1997) present a comprehensive statement of physical principles.

Two particularly important developments in our study of the Earth have resulted from paleomagnetism. We know how the geomagnetic field behaves on a billion-year time scale and we can trace the movements of surface features of the Earth, originally referred to as continental drift and now as plate tectonics. The crucial link is that, averaged over 10^4 years or more, the Earth's magnetic and geographic (rotational) axes coincide. This is known as the geocentric axial dipole (GAD) hypothesis. It follows that, with appropriate care in averaging, paleomagnetic measurements give the latitude and orientation (with respect to north) of a body of rock at the time of its formation. For rocks with ages exceeding a few tens of millions of years, the magnetizations are found to be incompatible with their present latitudes and orientations. Initially it was supposed that the observations could be explained by 'polar wander', that is, coherent motion of the whole of the Earth's surface relative to the rotation axis, but systematically different polar wander paths for different continents could be interpreted only by relative motion. Continental drift, postulated in the 1800s but considered seriously only by a small minority of Earth scientists until the early 1950s, rapidly became the conventional wisdom and a great scientific revolution was in progress.

Another vital discovery of paleomagnetism that became a cornerstone of plate tectonics is the sequence of polarity reversals of the magnetic field. At the ocean ridges, the upwelling centres of mantle convection, fresh igneous crust is produced, becoming magnetized as it cools. It then moves away from the ridges, producing stripes of alternating magnetic polarity in the oceanic crust that have a spacing characteristic of the reversal sequence. These allow the rate of sea floor spreading to be determined from marine magnetic surveys. Reversals themselves present interesting questions for dynamo theory. Why is the frequency of reversals so variable? Does the dipole field disappear during a reversal, leaving only the non-dipole field, or does an equatorial dipole remain? If so, why does the equatorial dipole appear to have preferred orientations?

25.2 Magnetic properties of minerals and rocks

Magnetic properties of materials are dominated by the intrinsic magnetic moments of electrons (electron spin). In a few materials the electron spins on neighbouring atoms interact to cause mutual alignment and, when large numbers of atoms have parallel magnetic moments, we say that the material is spontaneously magnetized. In bulk magnetic material, the regions of spontaneous magnetization, termed magnetic domains, are normally arranged in patterns to make the magnetic field follow closed loops, so that no total magnetic moment is observed. This is referred to as the demagnetized state. Application of an external field aligns the domains and the alignment may remain, wholly or partly, after the external field is removed. The material then has a magnetic remanence. The remanence that is induced in rocks by the Earth's magnetic field during their formation is, in many cases, stable enough to be measured millions or billions of years later. This is the basis of paleomagnetism.

There are several kinds of electron spin interaction, with different magnetic effects, illustrated in Fig. 25.1. The term ferromagnetism (iron-like magnetism) is loosely applied to the first, third and fourth of the patterns in the figure, all of which give spontaneous magnetizations. True ferromagnetism, in which all neighbouring magnetic moments are aligned parallel, is restricted to a few metals and alloys. It is observed in the metal constituents of meteorites and lunar samples, but not in terrestrial rocks. Strongly magnetic minerals are all of the ferrimagnetic type, in which neighbouring moments are aligned antiparallel, as in antiferromagnetism, but unequal numbers or strengths give a net magnetization. Ferrimagnetism refers to the magnetism of ferrites, which are oxides of iron and other metals that are of commercial interest as insulating magnetic materials. Magnetite (Fe_3O_4) is a ferrite. Its solid solutions with ulvospinel (Fe_2TiO_4), known as titanomagnetites, are the most important magnetic minerals in igneous rocks. They are sometimes found also in sediments produced from eroded igneous rocks.

The spontaneous alignment of electron spins in a ferromagnetic or ferrimagnetic material is an example of a cooperative phenomenon, which means that the probability of any one of them being aligned depends on the alignments of its neighbours and thus on the average of the

Interaction Type	Ferromagnetic	Antiferromagnetic	Ferrimagnetic	Canted antiferromagnetic
Examples	Fe, Co, Ni	NiO, MnO	Magnetite (Fe_3O_4)	Hematite (Fe_2O_3)
Atomic magnetic moments	↑↑↑↑↑↑	↓↑↓↑↓↑↓	↑↓↑↓↑↓↑	↗↘↗↘↗
Net spontaneous magnetization	↑	Zero	↑	→

FIGURE 25.1 The four most important patterns of mutual alignment of magnetic moments of atoms, caused by interactions of their electron spins.

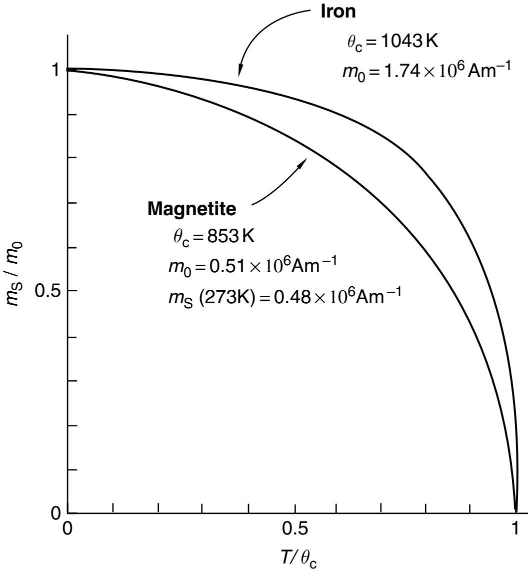


FIGURE 25.2 Temperature dependences of spontaneous magnetization, m_S, for magnetite and iron. Values are normalized to zero temperature values, m_0, and temperatures are normalized to the Curie points, θ_C.

whole assembly. As temperature is raised, the misaligning influence of thermal agitation increases, until a critical temperature is reached and spontaneous magnetization rapidly disappears. This is the Curie point, above which all these materials are paramagnetic, i.e. weakly magnetic, with no remanence possible. The variation of spontaneous magnetization with temperature for magnetite is compared with that for iron in Fig. 25.2.

The iron ions in magnetite are a mixture of Fe^{3+} and Fe^{2+} in the ratio 2:1, with the moments of the Fe^{3+} ions oppositely aligned so that the net spontaneous magnetization is that of the Fe^{2+} ions. When igneous rocks are weathered, the magnetite is oxidized to make all of the iron trivalent. The product may be maghemite (γFe_2O_3), which is a ferrimagnetic mineral structurally similar to magnetite. But maghemite is only metastable and converts to hematite (αFe_2O_3), which has a weak (parasitic) ferromagnetism due to the canting of its equal and nearly opposite atomic magnetic moments (Fig. 25.1). Hematite is the principal magnetic mineral in sedimentary rocks. Although only weakly magnetic, its remanence is often very stable. Hematite may also be formed, not directly or via maghemite, but via a hydrated oxide, goethite (FeOOH), which has properties similar to those of hematite. Also encountered, especially in sedimentary and metamorphic rocks, are ferrimagnetic iron sulphides, pyrrhotite and greigite.

Magnetic properties, such as the stability of remanence, depend on grain size, crystal imperfections, impurities and stresses. The smallest grains of any material are single domains, that is, they have the same direction of spontaneous magnetization throughout. They are permanently magnetized to saturation. The magnetic moment of a grain may be deflected by an applied field, but the only possible change in its remanence is from one preferred, or easy, direction of magnetization to another, usually the opposite direction. At low temperatures, reversal of the magnetic moment requires a high magnetic field and so single domains may have high magnetic stability, but, being small, they are subject to thermal agitation. This can cause repeated spontaneous reversal of their moments and the stability is then lost. This phenomenon is referred to as superparamagnetism. A material in which it occurs has a high magnetic susceptibility, that is it has a strong induced magnetization when exposed to a field, but cannot retain remanence. We refer to it below in connection with thermoremanence, the process by which superparamagnetic grain moments are stabilized by cooling.

Large grains are subdivided into domains, each of which is magnetized to saturation, geometrically arranged to form paths of magnetic flux closure. This structure avoids the appearance of surfaces of magnetic polarity either internally or on grain surfaces and so minimizes the magnetic energy. (For a review of ferromagnetic domain theory, see Kittel (1949).) The domain walls, separating domains with different directions of magnetization, have a progressive rotation of spin orientation and, in magnetite, are about 0.1 μm thick. This is a critical size for domain structures. Magnetite grains smaller than this cannot accommodate a domain wall. Spontaneous subdivision into domains occurs only for larger grains. The weak spontaneous magnetization of hematite allows much larger single domains. Changes in magnetization of multidomain grains occur mainly by movements of domain walls, enlarging domains that are favourably oriented with respect to an external field at the expense of their neighbours. This is an easier process than the coherent changes in direction of whole domains and so large multidomain grains tend to be magnetically soft and therefore of little interest to paleomagnetism.

For materials such as magnetite, true single domain grains are smaller than about 0.1 μm in size and true multidomains, with no observable single-domain-like properties, are larger than 10 μm to 20 μm. There is a wide intermediate range, for which grains are too large to be single domains, but nevertheless have properties that resemble them and are clearly incompatible with the theory of multidomains. We refer to them as pseudo-single-domain (PSD) grains. Most of the stable remanence that is of interest to paleomagnetism is attributed to them. So, an understanding of PSD properties is important to the subject, but they can be accounted for by a wide range of ferromagnetic domain structures and there is no single theory. We illustrate the problem considering the simple case of a grain with just two domains, with opposite magnetizations, separated by a domain wall, within which the electron spin alignments have a progressive rotation from one alignment to the other.

If the grain is perfectly symmetrical, with the two domains equal in size, then the grain has a magnetic moment perpendicular to them, arising from the magnetization within the domain wall. This domain wall moment can be reversed but not removed by any demagnetization process (except heating above the Curie point). Also, precise symmetry (equal domains) is unlikely to occur, because of both irregular grain shape and internal strains. Magnetization is accompanied by magnetostriction, the small change in crystal lattice spacing arising from an interaction of the electron spins with the electron orbits. This means that the domain wall involves strain energy and interacts with various crystal defects, which also cause strain energy. Minima in total energy occur for domain wall positions controlled by the distribution of imperfections. The wall may jump between them, a phenomenon well observed as Barkhausen noise in larger magnetic systems, in which magnetization changes are observed to occur in erratic jerks. Except in the unusual case of perfect symmetry there is a minimum component of the net moment in the direction of one or other of the domains. So, even in this simple case, we see that there are two components of PSD magnetization, arising from domain wall moments and Barkhausen discreteness.

Irregular grain surfaces and resulting complicated patterns of flux closure domains must also contribute to PSD moments. The important result of all these effects is that, unless the grains are so large that PSD effects are masked by much larger domains, they have minimum magnetic moments that respond to external fields in a manner similar to single domains. Most importantly, being larger than true single domains, they are less sensitive to thermal agitation. The most stable magnetizations are of the PSD type.

When a rock containing magnetic minerals is cooled in a magnetic field it becomes magnetized as soon as the temperature falls below the Curie point, θ_C. However, if the field is removed when the temperature is below but still close to θ_C the magnetization disappears. The domains are still spontaneously magnetized, but thermal agitation rearranges and realigns them, causing statistical cancellation in the case of small grains or rearrangement of domains to give flux closure in large grains. After cooling in a field to a blocking temperature, several tens of degrees below θ_C, magnetization may remain after removal of the field. This is thermoremanent magnetization, often abbreviated to its acronym, TRM. Each of the magnetic constituents of a rock, and perhaps even the individual domain walls, has a blocking temperature, below which thermal activation is ineffective and TRM becomes frozen in. Further cooling causes an increase in the TRM, according to the increase in spontaneous magnetization and, more importantly, an increase in its stability, that is, its ability to resist changes over a very long time. For small inducing fields the strength of thermoremanence is directly proportional to the field strength. Igneous rocks with natural remanent magnetization (NRM) that is thermoremanent in origin are favoured for determinations of the intensity of the ancient geomagnetic field.

The long-term stability of thermoremanence is crucial to paleomagnetism. It is most easily understood in terms of the single domain theory of Néel (1955). An isolated single domain of a material, such as magnetite, is anisotropic; its magnetic energy is a function of the direction of its magnetization. It is spontaneously magnetized in one of the so-called easy directions, between which there are energy barriers that can be overcome either by application of a strong field or by thermal activation. Anisotropy arises from crystalline effects and also from grain shape. Unless a grain is almost equidimensional, the shape effect is stronger and has the effect of making the longest axis the easy direction, because that minimizes the external energy of its field. The two opposite directions are equivalent and, if an energy barrier, E, separates them, then the probability, dP, that a thermal impulse will reverse the magnetization in time dt is

$$\mathrm{d}P = \nu_0 \exp(-E/kT)\mathrm{d}t, \tag{25.1}$$

where ν_0, termed the attempt frequency, is of order 10^9 Hz.

Now consider a single domain grain that is cooling from a high temperature, with the thermally activated reversals of its magnetization becoming progressively more sluggish. If the cooling rate is represented by a characteristic time, τ, for cooling through a few degrees, which would be ~100 s in a laboratory experiment, then the blocking temperature, T_{B}, is the value of T in Eq. (25.1) at which $\tau(\mathrm{d}P/\mathrm{d}t) \approx 1$. Thus we can write

$$E/kT_{\mathrm{B}} = \ln(\nu_0\tau) \approx 25.3. \tag{25.2}$$

Further cooling has two effects. (E/kT) increases with decreasing T, but there is also an increase in E, which, for the case of shape anisotropy, is proportional to the square of the spontaneous magnetization. By ignoring the variation in E we overestimate dP/dt at any lower temperature, T_{L}, that is

$$\mathrm{d}P/\mathrm{d}t < \nu_0 \exp(-25.3T_{\mathrm{B}}/T_{\mathrm{L}}). \tag{25.3}$$

We are interested in high blocking temperatures, which give greatest stability. Assuming $T_{\mathrm{B}} = 750$ K, 100 K below the Curie point of magnetite, and $T_{\mathrm{L}} = 300$ K, Eq. (25.3) gives $\mathrm{d}P/\mathrm{d}t < 3.4 \times 10^{-19}$ Hz, which corresponds to a relaxation time exceeding 9×10^{10} years.

We see how it is possible for magnetic remanence to be stable over geological time, but even rocks with high magnetic stability include components with lower stability. The less stable components may gradually acquire magnetizations subsequent to rock formation, at times when the field was different, a phenomenon termed magnetic viscosity. Viscous magnetization may be enhanced by mild reheating that is insufficient to affect the primary remanence, but introduces a disturbing secondary magnetization. Secondary magnetizations are softer than the primary remanence and their removal by partial demagnetization in an alternating field (in zero ambient field), 'AF cleaning', is a standard method of revealing the primary remanence. Partial thermal demagnetization is also useful as it retraces, in reverse, the acquisition of thermoremanence, allowing identification of the blocking temperatures of different components. Heating experiments of this kind require cross-checks to discard observations that are affected by chemical changes.

Above their blocking temperatures, single domains are superparamagnetic. For a magnetic moment μ in a field B, alignment with B is favoured by the Boltzmann factor $\exp(\mu B/kT)$, or $\exp(-\mu B/kT)$ for the opposite alignment. The average magnetization is proportional to the difference between these factors. Relative to the saturation magnetization, m_{s} (all moments aligned parallel), this is

$$\begin{aligned}\frac{m}{m_s} &= \frac{\exp(\mu B/kT) - \exp(-\mu B/kT)}{\exp(\mu B/kT) + \exp(-\mu B/kT)}\\ &= \tanh\left(\frac{\mu B}{kT}\right).\end{aligned} \tag{25.4}$$

P. Langevin generalized this expression to describe ordinary paramagnetism, for which μ is very much smaller, being the value for a single atom, and the magnetization is correspondingly weaker. The Langevin expression is simply an integral over all orientations of grain moments with respect to the field. Hence the term 'superparamagnetism' for the behaviour of spontaneously magnetized but thermally randomized single domains. Cooling through the blocking temperature freezes in the superparamagnetic moment, which becomes thermoremanence. The only change with further cooling is an increase in remanence in proportion to the increasing spontaneous magnetizations of the domains.

Equation (25.4) would apply to an assembly of single domains with their axes parallel to the inducing field, but is readily integrated over a randomly oriented distribution, as in the Langevin theory. The tanh form is characteristic of single domain behaviour. For small fields ($\mu B \ll kT$), thermoremanence is proportional to the inducing field, but it begins to saturate in much lower fields than does multidomain remanence. In spite of the fact that true single domains are probably rare in rocks, the single domain theory is of greater interest than multidomain theory because it is a reasonable description also for pseudo-single domains, which are the carriers of the most stable remanence.

When an igneous rock is eroded and its fragments are deposited in water, the magnetic grains may carry remanent moments from their earlier history and, if the water is very still, they may be partly aligned by the ambient field at the site of deposition. The resulting sediment acquires detrital remanent magnetization (DRM), but true DRM is generally transient. The subsequent realignment of grains during sediment consolidation is accompanied by a stronger post-depositional remanence. This, too, is lost if there are chemical changes. Then new magnetic minerals are formed, that may be hematite in an oxidizing environment or pyrrhotite in a reducing environment with available sulphur. They acquire chemical remanent magnetization (CRM), the remanence resulting from chemical formation in a field. This is normal in sedimentary rocks. Also, in some igneous rocks, such as sea-floor basalt with prolonged water percolation, the remanence may be more chemical than thermoremanent in origin. CRM is similar in stability to TRM and equally useful for paleomagnetic direction studies, provided the time of its acquisition is reasonably well known, but not for paleointensities (Section 25.5).

25.3 Secular variation and the axial dipole hypothesis

Although the natural magnetizations of rocks and archeological samples record both the directions and intensities of the fields in which they were formed, the directions can be determined more easily and more reliably than the intensities. If a large fraction of a primary remanence survives the process of magnetic cleaning to remove secondary magnetizations, then the primary direction is accurately determined. The interpretation of intensity is less straightforward. Additional experiments, involving heating, are required and chemical changes caused by heat make the interpretation both more difficult and more prone to error. For this reason, studies of the intensity of the ancient field have proceeded more or less independently of the directional studies. This section is concerned with the directional information and paleointensities are considered in Section 25.5.

Magnetic field directions obtained from the remanence in pottery kilns for which the last firings have been carbon-dated are plotted in Fig. 25.3. This extends the historical record of the secular variation in Britain to about 2000 years. It is a plot of magnetic inclination vs declination, known as a Bauer plot, after L. A. Bauer, who first represented secular variation in this way. Such a graph is a series of spot measurements of the field in a particular area. It is satisfactory to compare in this way observations made over distances up to about 1000 km, because the strong features of the field are larger than this (Fig. 24.1). The record has been extended back to 10 000 years, although with slightly less precision, using thin slices of lake sediment (Turner and Thompson, 1981; Creer and Tucholka, 1982), and shows a similar character for the whole period. If the field were due only to an axial dipole at the centre of the Earth

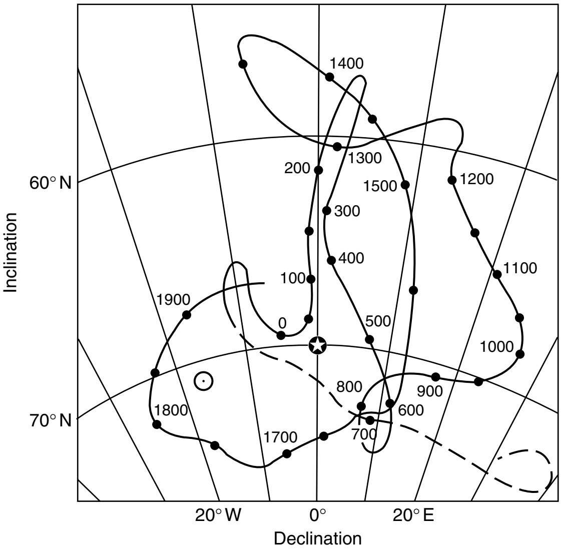


FIGURE 25.3 Secular variation of the magnetic field in Britain, plotted as inclination vs declination as a function of time. The record of direct observations from about 1600 is extended back by archeomagnetic measurements, with dates given by numbers on the curves. Archeomagnetic data are from a plot by Tarling (1989). The direction of the field of a geocentric axial dipole is shown by the star and the direction corresponding to the present, inclined dipole is represented by the open circle.

then the direction would be that of the star in the centre of the figure. The figure shows, and more extended data confirm that, although the field drifts about, it is constrained not to wander too far from that direction and the wandering is centred on it. The present, tilted dipole field gives the direction indicated by the open circle, which is clearly displaced from the centre of the figure.

Another way to represent paleomagnetic directions is to plot the virtual or apparent pole positions, assuming the observed field to be due to a dipole. If each of the spot measurements used to produce Fig. 25.3 had, instead, been used to calculate a virtual pole position, then the virtual magnetic pole would have traced out a similar-looking path. But there would be a subtle difference, because the pole position is calculated from inclination by Eq. (24.11), which is not linear. Thus, the mean of the independently determined virtual pole positions would not coincide precisely with the pole obtained from the mean direction in Fig. 25.3. Differences of this kind have become important to the fine details of the paleomagnetic record and will be considered further presently.

As data sets such as that in Fig. 25.3 are extended to longer periods, so it becomes even clearer that the average field direction is that of an axial dipole, not just for a single site, but for all sites. There is a limit to the time span that can be considered in this way because the slow processes of polar wander and continental drift (Section 25.5) eventually spread the paleo-poles around the globe. However, even up to 20 million years ago, the scatter of virtual pole positions appears centred on the present geographic pole. Figure 25.4 is a plot of pole positions obtained from igneous rocks younger than 20 million years. The points are spot measurements of the field directions for the brief periods when the rocks were cooling. Although they were obtained from many places, and so are not points on a single virtual pole path, they are from many similar paths and the effect is the same. For the purpose of plotting Fig. 25.4, the polarity of the field is ignored. It has reversed many times in the past 20 million years and whichever pole happened to be in the northern hemisphere is plotted. Reversals are considered in Section 25.4.

The conclusion that the averaged geomagnetic field is that of a dipole at the centre of the Earth, with its axis coinciding with the geographic axis, is an important fundamental result in paleomagnetism. It is of interest in dynamo theory, but its great impact has been in establishing continental drift. Suitably averaged paleomagnetic poles are not just average magnetic poles, they are also geographic poles. Thus the wander of the magnetic pole, viewed over hundreds of millions of years, traces the rotation axis, relative to surface features where paleomagnetic samples are obtained. This application is pursued in Section 25.6.

With close examination of observations, the axial dipole hypothesis is seen as a good approximation, but not exact. Curie's principle of symmetry, which we refer to in Section 24.5, gives confidence to the expectation that, averaged over a sufficient time, the field is symmetrical about the rotation axis. But this allows an axial

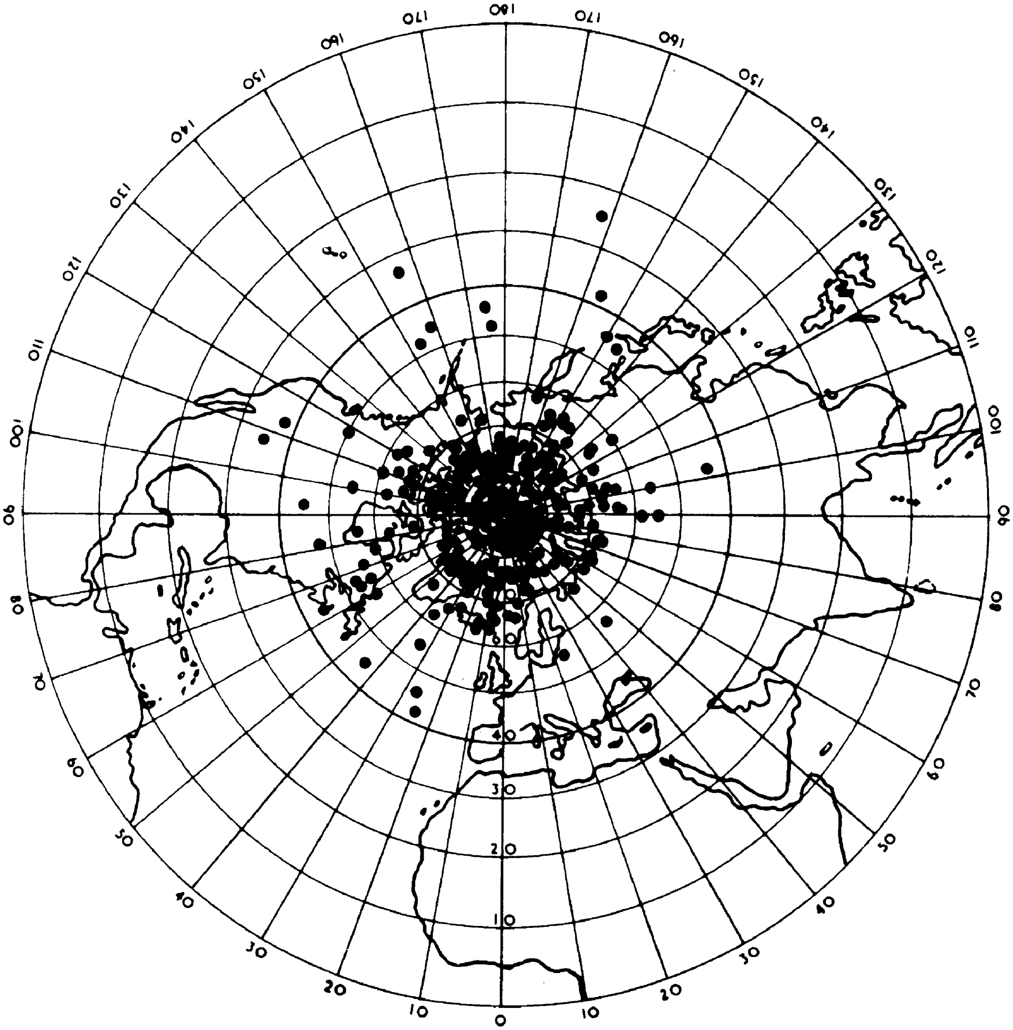


FIGURE 25.4 Paleomagnetic pole positions obtained from igneous rocks up to 20 million years old. Reproduced, by permission, from Tarling (1971).

quadrupole and higher terms, as long as they are axial, and does not even demand a dominant dipole. Remoteness of our observations from the source of the field makes a spatial spectrum with the general form of Fig. 24.2 almost inevitable, so from the perspective of paleomagnetism the axial dipole hypothesis looks safe, but we have a reason for enquiring how good an approximation it is, apart from the fact that we rely on it to interpret paleomagnetic data.

Theories agree that core motion within the tangent cylinder, parallel to the rotation axis and enclosing the inner core, is semi-isolated from the convective motion in the rest of the outer core. If this has a surface expression in the pattern of the field, it would be most obvious as an axial quadrupole, systematically related to the dipole. The present field does not provide a sufficient test for a systematic quadrupole and statistics of a much longer record are needed to examine this possibility. From a study of paleomagnetic data for the last five million years, Constable and Parker (1988) concluded that there is a consistent axial quadrupole, represented by a harmonic coefficient g_2^0 that is about 6% of the axial dipole term, g_1^0, and

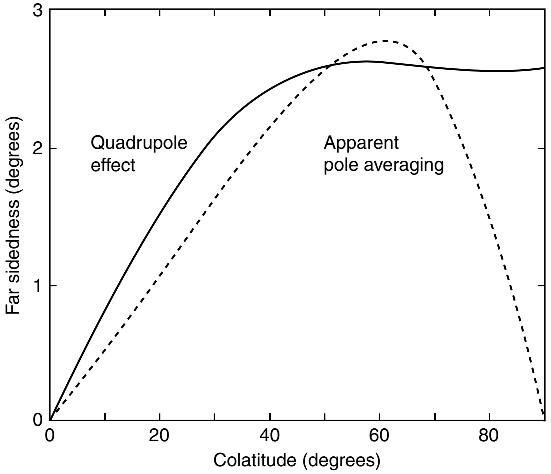


FIGURE 25.5 Apparent angular displacement of the pole (far-sidedness) as a function of latitude due to a 6% quadrupole field by Eq. (25.5) (solid line), compared with the effect of taking averages of magnetic inclination for a dipole field transiently deflected to ±15° from the true pole (broken line). This illustrates the need for care in statistical averaging of secular variation if small effects, such as the quadrupole field, are to be discerned.

reverses with the dipole. Merrill *et al.* (1996, Table 6.2) give 3.8%. This is evidently a robust result, in spite of the difference between these estimates, and their dependence on assumptions about the statistical behaviour of the field, a point that we return to below. Selecting only the g_1^0 and g_2^0 terms in Eq. (24.14), differentiating to obtain the radial and circumferential components of the field by Eqs. (24.19) and (24.17) and taking the ratio, the dip angle is given by

$$\tan I = \frac{B_r}{B_\theta} = \frac{2g_1^0 \cos\theta + 3g_2^0 \cos\theta\left(\frac{3}{2}\cos^2\theta - \frac{1}{2}\right)}{g_1^0 \sin\theta + 3g_2^0 \sin\theta\cos\theta}. \tag{25.5}$$

The apparent colatitude inferred by supposing that this magnetic inclination is due to the axial dipole alone is obtained by Eq. (24.11). The difference between this apparent colatitude and the true colatitude is plotted in Fig. 25.5 for the ratio $g_2^0/g_1^0 = 0.06$, as estimated by Constable and Parker (1988).

Figure 25.5 is a plot of what Wilson (1970) was first to identify as a systematic 'far-sidedness' of paleomagnetic poles. The magnetically inferred colatitude exceeds the true colatitude by a small angle, so that, from all directions, the pole appears to be slightly further away than it really is. Of course, such conclusions are possible only by analysing data from rocks that are young enough for polar wander/continental drift to be insignificant. As Fig. 25.5 shows, the apparent polar displacement is sensibly constant over most of the Earth's surface. With the 6% assumption it is between 2.0° and 2.6° for the colatitude range 28° to 90°, which is almost 90% of the area and excludes only the little sampled polar regions. The far-sidedness has no obvious latitude dependence. For it to be a systematic effect it is necessary for g_2^0 and g_1^0 to have the same sign, reversing together, and for the other harmonics to be averaged out by sufficient sampling to allow a 2° deflection to be recognized as statistically significant. The scatter of paleomagnetic directions is much greater than this so the procedure for averaging the secular variation requires close scrutiny.

A basic problem is that a pole position is calculated from magnetic inclination by Eq. (24.11), which is non-linear. Merrill *et al.* (1996, Section 6.4) discuss the statistical arguments that have been used to deal with this problem. A random distribution of field directions about a mean does not produce a random distribution of inferred poles or vice versa. To illustrate this point, consider a site where the magnetic dip angle corresponding to an axial dipole is 45°, but a pair of measurements give 30° and 60°, representing symmetrical scatter, arising from local, non-dipole field variations. The colatitude is $\theta = \mathrm{ctn}^{-1}(½ \tan 45°) = 63.4°$, but, assuming a dipole field in each case, the two observations correspond to 73.9° and 49.1° and so to an average of 61.5°. This bias would indicate near-sidedness and so give an underestimate of observed far-sidedness. Conversely, if the scatter of field directions is due to a randomly wandering pole, that would produce a spurious far-sided effect. Magnetizations at colatitude 45° due to poles at 30° and 60° would have inclinations of 73.9° and 49.1°, averaging 61.5° and indicating a pole at 47.4°, far-sided by 2.4°. But, if measurements are made on a sediment that does the averaging, then it is the vector average field that is observed and not a simple average of directions. Assuming a similar dipole strength for both pole positions, by Eq. (24.10) the nearer pole gives a field of $1.8B_0$ and the more remote one a field of $1.3B_0$, biasing the average direction to the nearer pole and giving an average inclination of 63.4°. Since this corresponds to the true latitude of 45°, the sedimentary weighted average compensates for the bias introduced by the non-linearity of Eq. (24.11).

We may suppose that drift of the non-dipole field past a site would cause angular dispersion of the field direction that is properly averaged at the site before calculation of the pole position, but that components of the secular variation caused by the equatorial dipole, and variations in the strength of the axial dipole, require averaging of the pole positions corresponding to spot measurements of the field. This suggests a decrease in the scatter of apparent pole positions with the latitude of observation, because the dipole field is stronger at high latitudes, reducing the scattering effect of the non-dipole field. However, the opposite latitude variation is observed, so it is evident that we cannot separate dipole and non-dipole effects in this way.

McFadden *et al.* (1988) argued that a more useful procedure is to consider separately the symmetric and antisymmetric harmonic components of the field. This refers to symmetry about the equator. The axial dipole, represented by g_1^0, is antisymmetric, but the axial quadrupole, g_2^0, is symmetric. These are the leading members of two 'families' of harmonics, which have been referred to as the dipole and quadrupole types, but this is misleading because g_1^1 and h_1^1, which represent the equatorial dipole, are symmetric and belong to the quadrupole family with g_2^0. In terms of harmonic order, l, and degree, m (see Appendix C), the distinction is between an antisymmetric family, for which $(l-m)$ is odd, and a symmetric family with even $(l-m)$. From an examination of the harmonic components of the present field, McFadden *et al.* (1988) observed that the scatter of directions from an axial dipole field attributable to the symmetric components is independent of latitude, but that antisymmetric components cause scatter approximately proportional to latitude. The combination gives the observed increase with latitude. The conclusion is that correction for paleomagnetic scatter should involve two terms, one constant and one proportional to the apparent latitude of a measurement. While this approach allows a more rigorous assessment of the significance of the quadrupole field, it does not completely solve the problem of paleomagnetic scatter. But we can note that it involves an error of 2° to 3° at most, and that this is very small compared with the angular variations observed in paleomagnetism.

This discussion prompts a reconsideration of the distinction between odd and even harmonics. It suggests that, for the purpose of dynamo theory, the fundamental distinction is between odd or even $(l-m)$, rather than odd or even l. Merrill *et al.* (1996) refer to a discussion by P. H. Roberts and M. Stix of α–ω dynamos of the Bullard and Gellman type (Section 24.5), pointing out that, if the core velocity field is symmetrical about the equator, then the ω-effect, caused by differential rotation, is also symmetrical, but that the α-effect, which depends on helicity, is opposite in the two hemispheres. In this situation, the odd and even families of harmonics may be generated independently, interacting only to the extent that the pattern of core motion is asymmetrical. Even if core motion is too irregular for such a clear separation, a difference between the within-family and between-family interactions must have a fundamental implication for the behaviour of the field. This argument is considered in the following section in connection with reversals. The separation of the harmonic components of the field into antisymmetric and symmetric families refers to the observed, poloidal field.

Merrill *et al.* (1996) discount the significance of an axial octupole term, g_3^0, but Kent and Smethurst (1998) reported evidence that paleomagnetic data from the period before 250 million years ago suggest $g_3^0/g_1^0 \approx 0.25$. Unlike the situation in the last five million years, the continents have been redistributed since (and during) that time so it is not possible (without a much larger coherent continent than actually exists) to isolate such an effect from polar wander/ continental drift. Only a statistical argument is possible. The Kent and Smethurst evidence is a strong bias to shallow magnetic inclinations, which would be consistent with a random distribution of continents if there were such a strong octupole field. The alternative interpretation is a clustering of the sampled continents at low latitudes, which Pesonen *et al.* (2003) found to be the case in a series of Proterozoic continental reconstructions. There is no evidence that the field was so fundamentally different in the early period that an octupole field much stronger than the dipole field at core level is plausible. Rather, we would expect the dipole field to have been relatively stronger when the inner core was small. So, the question raised by the Pesonen *et al.* (2003) analysis is: what reason might there be for continents to cluster around the equator? If we had an answer it would make an interesting topic for Chapter 12.

Over-riding the question of the dipole/quadrupole or multipole structure of the field is its axial character. By Curie's principle of symmetry, which we refer to also in Sections 17.9 and 24.5, when averaged over a suitable time, the magnetic field must be symmetrical about the rotation axis, unless there is some permanent asymmetry in its causes. Since we are not aware of such an asymmetry, and it seems unlikely that there is one, we accept the axial principle, which is central to paleomagnetism and to global tectonics (Section 25.6).

25.4 Geomagnetic reversals

The discovery of thermoremanent magnetization (TRM), induced in pottery, bricks and lavas as they cooled, dates from at least the early nineteenth century. Certainly it was well known in 1906 when the French physicist B. Brunhes found not only a lava flow but also an adjacent clay that had been baked by the lava, in which the remanent magnetism was almost precisely opposite to the present direction of the field. Brunhes drew the conclusion that the Earth's field must have reversed, but his discovery was so far in advance of dynamo theory that there was no way that it could be appreciated fully at the time. Further similar observations in the 1920s by P. L. Mercanton and M. Matuyama also attracted little attention until much later.

By the late 1940s, many more reversely magnetized rocks were being discovered, but the concept of a reversed field was not accepted without question and alternative explanations were sought. L. Néel, who had developed the theory of ferrimagnetism (see Fig. 25.1), undertook an exhaustive theoretical study of mechanisms by which magnetic remanence in minerals could, in principle, be self-reversing. Not all of Néel's mechanisms were really plausible, but in 1952 a lava from Mt Haruna, in Japan, was found to acquire reversed TRM in laboratory experiments. An intensive period of theoretical and laboratory studies established that the self reversal was due to ionic rearrangement in solid solutions of ilmenite ($FeTiO_3$) and hematite (Fe_2O_3) (Ishikawa and Syono, 1963). The disordered, high temperature state converts to an ordered arrangement of ions at low temperatures. Ions that move to new lattice sites are magnetically aligned by interactions with their neighbours and develop a reversed remanence exceeding the original normal one. The reversed remanence is a property only of an intermediate, partially ordered state and is not observed in either the disordered or fully ordered states. It is a rare phenomenon, but the fact that it occurs allowed expressions of doubt about geomagnetic reversals to persist for several years.

Studies dedicated to the reversal problem eventually made the case for field reversals unassailable. There are four principal observations.

(i) Wilson (1962) found numerous baked contact rocks, of the type first recognized by Brunhes, and showed that in the overwhelming majority of cases their remanence directions coincided with those of the lavas that heated them and generally not with the directions in neighbouring unheated rock. Later statistics with larger numbers have reinforced this conclusion.

(ii) For rocks up to a few million years old, which can be sufficiently accurately dated, the dates of normal and reversed rocks coincide on different continents and for different rock types (e.g. lavas and sediments). By convention, magnetizations parallel to the present field are referred to as 'normal', although it is clear that there is nothing abnormal about reversed magnetization.

(iii) The detailed process of reversal has been traced in both rapid sequences of lava flows and in deep sea sediments.

(iv) Linear magnetic anomalies observed at sea (as in Fig. 12.10) are identified with stripes of normal and reversely magnetized basalt, parallel to the ocean ridges where they were formed (Fig. 12.9). For the few million years of dating that is sufficiently accurate for a precise check, the sequence of reversals required to explain the anomalies coincides with the magnetic polarities of dated igneous rocks, assuming a more or less steady spreading of the sea floor away from the ridges. This is consistent with paleontological estimates of the age of the ocean floor (Fig. 12.8).

Dynamo theory readily accommodates reversals, because a field of either sign can be maintained equally effectively by any particular pattern of motion or body forces. Equation (24.36) is unaffected by reversing the sign of **B**. All that is needed is an instability that triggers reversals. The secular variation demonstrates that there is no steady, stable state of the dynamo, but reversals do not appear to be just extremes of the normal secular variation, although the two phenomena are presumably related. Gubbins (1994) reviewed the basic physics of this problem. Growth, decay and drift of the non-dipole field and of the equatorial dipole are continuous processes, and the axial dipole also fluctuates in strength, but reversals of the axial dipole occur in 5000 years, or perhaps less in some cases, compared with the average interval between them, hundreds of thousands of years. This pattern of behaviour is suggestive of a chaotic system switching irregularly between two quasi-stable states. However, there are extreme variations in the frequency of reversals that require an explanation.

We can examine details of the record of reversals for clues to the dynamo instability that causes them. There are two semi-independent aspects to the problem, the field pattern during a reversal, and the statistics of the intervals between reversals. There is wider general agreement on the statistical behaviour. The normal and reversed states occur equally often, with no difference between their average durations, as would be expected for a purely random process that is the same for both polarities. If we suppose that reversals are independent random events, with a fixed average rate of occurrence, then the probability $p(t)$ of an interval of constant polarity with duration t decreases exponentially with t (in the same way as the probability of survival of a radioactive nucleus). This is the Poisson distribution

$$p(t) \propto \mathrm{e}^{-t/\bar{t}}, \tag{25.6}$$

where $\bar{t}$ is the average interval of fixed polarity between reversals and is the same for both polarities. However, $\bar{t}$ is not constant with time, but has varied slowly from about 2×10^5 years to more than 10^7 years. Merrill *et al.* (1996) considered the possibility that a reversal might be followed by a period of inhibition to further reversals, requiring replacement of Eq. (25.6) by a gamma function. However, they noted that, with successive improved reversal chronologies, the departure from Eq. (25.6) diminished. It appears likely that there is no inhibition and that the statistical indication of it arises from short polarity events that are missed in the record (but see Problem 25.5).

Figure 25.6 shows the reversal record for the last six million years, for which the dating of reversals is most precise, and therefore the agreement between data from different continents, and between sediments and igneous rocks, is most securely observed. Back to about 100 million years ago the framework of the record is obtained from the sea floor magnetic anomalies. For earlier times, without the sea floor record, the sequence is less certain, although it is evident that the reversing behaviour was similar. Extended periods with one dominant polarity are referred to as chrons, with recent ones named after early pioneers in geomagnetism. Briefer polarity intervals (subchrons) are named after the sites where rocks that record them were first found. Even shorter

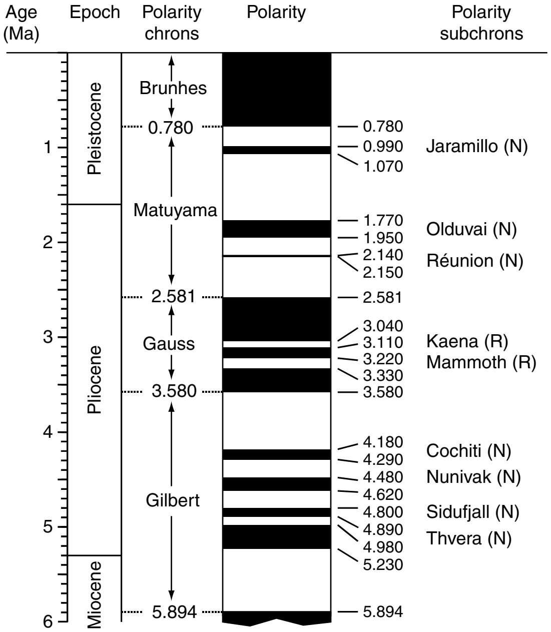


FIGURE 25.6 Geomagnetic polarity record for the last six million years, as plotted by Merrill *et al.* (1996) from data by Cande and Kent (1995). Periods of normal polarity are marked black and reversed polarity white. Numbers give dates of boundaries in millions of years.

events, not represented in the figure, occur and there are also incomplete or aborted reversals, termed excursions, that may be too brief to indicate that a full 180° polarity change occurred, or, if it did so at one site, whether that was apparent elsewhere. The sequence of reversals, with irregular long and short polarity intervals, can be seen in sufficient detail in sediment cores to allow dates to be correlated with greater precision than the determination of absolute ages. The use of the reversal record for stratigraphic correlation has become an important adjunct to paleontological and isotopic methods (Opdyke and Channell, 1996).

Detailed observations of field changes during a reversal require a continuous record, such as that in Fig. 25.7. This was obtained from a marine sediment core that had been deposited rapidly enough to minimize direction averaging. It shows several features that are common to reversal records, most significantly the diminished field strength for a few thousand years during the reversal and the progressive decrease for

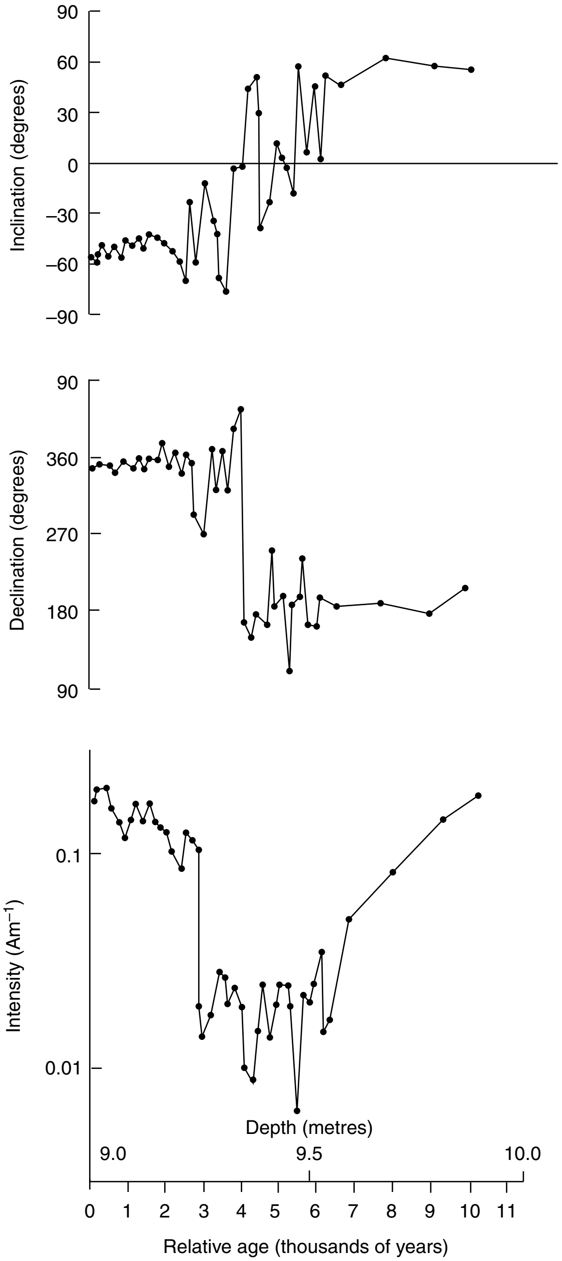
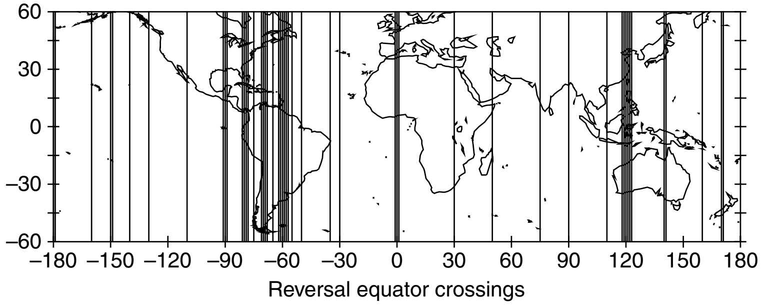


FIGURE 25.7 Detailed record of a reversal from a rapidly deposited deep-sea sediment core. Samples were 'cleaned' in a 10^{-2} T alternating field before measurement. This reversal marks the lower boundary of the Jaramillo polarity event, 1.07 million years ago (see Fig. 25.6). Note that time runs from right to left. Figure redrawn after Opdyke *et al.* (1973).

FIGURE 25.8 Longitudes of equator crossings of poles during polarity transitions observed by sedimentary cores, as plotted by McFadden and Merrill (1995).



several thousand years before there were significant direction changes. The inclination record shows repeated strong fluctuations, but this is probably not very significant because, with the weakened field, the non-dipole field was more prominent, causing direction changes that can be identified only with a particular locality and not with the polarity of the field as a whole.

Measurements of paleo-intensity during numerous reversals, such as that in Fig. 25.7, agree that the field strength is reduced by a factor of 3 to 10 and, as in this example, the reduced intensity is generally more prolonged than the period of rapid direction change. Since the dipole field normally dominates the observed field only by a similar factor, the obvious conclusion is that the axial dipole dies away, and redevelops in the opposite direction, with no accompanying enhancement of the equatorial dipole, or of the non-dipole field. An alternative, but disputed, suggestion is that the equatorial dipole remains strong enough for the field to retain its dipole character during a reversal, so that the magnetic north pole crosses the equator at an identified longitude, instead of disappearing from one hemisphere and reappearing in the other. The particular interest in this question arises from reports that the pole not only retains its identity, but that, during repeated reversals, it follows preferred paths from pole to pole, either through the Americas or 180° away, through East Asia and Western Australia. Laj *et al.* (1992) summarized the results of several authors who reached this conclusion. The interpretation is not entirely straightforward and, in reviewing the evidence, Merrill and McFadden (1999) questioned the significance of the preferred paths. We take the position that the effect is real and that the explanation is a mantle control on magnetic flux at the top of the core, for reasons that follow.

Figure 25.8 shows the representation by McFadden and Merrill (1995) of the longitudes of the reported equator crossings by the postulated reversing dipole. Note that it is the path of the geomagnetic north pole that is plotted, so that this is a polar effect for which the opposite polarities are not equivalent. Gubbins (2003) drew attention to the fact that the preferred paths provide the most direct indication of mantle control on core behaviour and presented further evidence. He plotted a graph of the square of the vertical component of the present field, averaged over all latitudes, as a function of longitude, and showed that it is highly correlated with a similar graph of shear wave speed at the base of the mantle. A strong vertical component (of either sign) is seen where the mantle appears cool. Gubbins's point is that cool mantle enhances the flux of core heat into it and the resulting cooled core material is carried down convectively. By the frozen flux principle it carries the field with it, towards the downwelling, and the surface expression is a locally strong vertical field. The striking feature of the Gubbins graphs is that there are two peaks, at longitudes coinciding with the preferred pole paths during reversals. This partially resolves what has been a paradox in the reversals story: the preferred paths can be explained simply as an artifact of mantle control that concentrates the vertical field component at selected longitudes.

However, a crucial question remains. The reported reversal paths show a preference for the

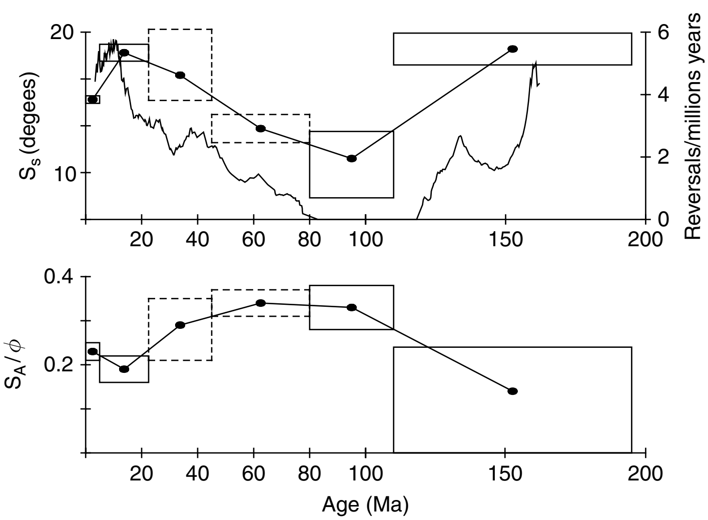


FIGURE 25.9 (Top) Variation of reversal rate for the last 165 million years (jagged line – right-hand scale), with the latitude-independent component of the scatter of pole positions due to secular variation during periods between reversals. (Bottom) Latitude-dependent component of the secular variation (see Eq. (25.7)). Based on McFadden *et al.* (1991) and redrawn from a figure by Merrill *et al.* (1996).

western path (through the Americas). Assuming this to be statistically significant, since the path of the magnetic north pole is plotted, conditions at the core–mantle boundary must distinguish between the two polarities. It is possible, in principle, for chemical differences to generate thermoelectric or galvanic currents that would bias the dynamo, although we believe the core–mantle boundary to be too nearly isothermal to admit the thermoelectric alternative as likely. The only serious possibility appears to be a chemical battery effect, varying with time as the material at the base of the mantle is replaced. This allows the suggestion that the seismic velocity variations at the base of the mantle, used in the analysis by Gubbins referred to above, have an origin at least partly in chemical differences. While this is just possible, we need to question the statistical significance of the polarity bias that it aims to explain.

Reversals occur sufficiently often to allow observation of a statistically significant variation in their rate of occurrence (Fig. 25.9). The peak rate is at least five per million years but, as the figure shows, there was an extended period of constant (normal) polarity (disallowing possible unobserved brief events) from about 118 million years to 84 million years ago (the Cretaceous superchron). An even longer period of reversed polarity (the Permo-Carboniferous superchron) appears to have extended from 312 to 256 million years ago, but is not confirmed by the sea floor anomaly record because no ocean floor has survived from that time. This figure documents the idea, discussed by McFadden *et al.* (1991), that reversals are most frequent when the symmetric harmonics of the field, referred to in Sections 24.3 and 25.3, are strongest, relative to the antisymmetric harmonics. They noted that the scatter of virtual poles caused by symmetric components of the present field, S_S, is independent of latitude, ϕ, but that the scatter by antisymmetric components, S_A, is proportional to ϕ. Spherical harmonics are orthogonal functions so that the total scatter, S, is given by

$$S^2 = S_A^2 + S_S^2 = (S_A/\phi)^2\phi^2 + S_S^2. \qquad (25.7)$$

By assuming this to be true also in the remote past, McFadden *et al.* separated the antisymmetric and symmetric components of the scatter, as in Fig. 25.9. Since we must suppose that the scatter is proportional to strengths of the components, this gives a measure of their relative strengths. The positive correlation of S_S (six averaged data points) with the rate of reversals (solid line) is clear in the top half of the figure. The inverse correlation with the antisymmetric field is shown in the bottom half.

While we grope for an understanding of the reversal mechanism, and hypotheses abound, we can sift the plausible from the implausible to narrow the range. Mantle control is sufficiently strongly indicated to consider the three obvious ways in which it may influence core motions: thermal, topographic and electromagnetic. They are not necessarily independent, but the extreme variability of the reversal rate implies a variation in core–mantle conditions that is localized and cannot reasonably be global. For example, the efficiency argument in Section 22.7 requires that any variation in the total core-to-mantle heat flux be reflected in dynamo power. The evidence is conflicting. Shaw and Sherwood (1991) reported no correlation between field strength and the rate of reversals, but Tauxe (2006) found field intensity and polarity interval duration to be positively correlated. Figure 25.9 indicates that the symmetric/antisymmetric ratio is diagnostic; it is possibly the most direct clue that we have. The reversal mechanism seen in the numerical dynamo model of Takahashi *et al.* (2005) may also offer a lead. In this model, concentrated radial flux patches appeared at low latitudes and their migration to high latitudes was followed by field reversal in a manner qualitatively similar to the equatorward migration of sunspots in the 11-year cycle of solar field reversals. Radial flux at the equator is characteristic of symmetric field components (axial quadrupole and equatorial dipole), so, if the core–mantle boundary structure is such as to promote the development of flux patches of the kind seen in the Takahashi *et al.* model, it could be reflected in the strength of the symmetric harmonic terms.

It may be possible to merge the flux patch idea with a theory by Olson (1983), which requires core motions to be driven by two sources of buoyancy (or negative buoyancy), one originating at the inner core boundary and the other at the core–mantle boundary. The idea is that the two sources of buoyancy generate motions of opposite helicities in the rotating fluid. The sign of helicity can be thought of as the direction of rotation as one follows a spiral path and is opposite in the two hemispheres for each source of buoyancy. The heterogeneity of the D″ layer at the base of the mantle allows the possibility of local variations in the heat flux conducted into it and, therefore, in the generation of localized negative thermal buoyancy at the top of the core.

Several, but not all, studies have indicated that the inner core inhibits reversals, because, being solid, the only way that the field in it can change is by diffusion. Its electromagnetic relaxation time is about 1600 years, which is shorter than the time for a reversal, but sufficient to add some inertia to the dynamo. If this is significant then we might expect reversals to have been more frequent when the inner core was smaller, but the detailed reversal history is too short to give an indication of the behaviour 2 billion years or so ago and, in any case, such an effect would probably not be observable against the background of variations that occur anyway. There is a better chance of seeing an effect of the change in inner core size from measurements of the intensity of the early field, because a smaller inner core contributes less of the two most important components of core energy – compositional separation and latent heat (see Section 22.7).

25.5 Paleointensity – the strength of the ancient field

There is a long history of paleointensity measurements, but reliability is a serious problem for several reasons. The now classical method was developed by Thellier and Thellier (1959), essentially by progressively heating and cooling igneous rock samples and pottery fragments to retrace the acquisition of their thermoremanence. The Thelliers experimented with partial thermoremanence (pTRM), which is induced in a sample that is cooled in a field through a limited temperature range and in zero field for the remaining temperature intervals. pTRM is identified with grains or domains that have blocking temperatures within the range of the field exposure. The experiments established the principle of additivity of pTRMs: the pTRMs acquired over all of the separate temperature ranges add to the total TRM observed if the sample is cooled in the field over the whole temperature range. The pTRMs acquired in a low field (the Earth's field is low in this context) are independent and each of them is lost by heating through the temperature range over which it was acquired. The

Thellier method compares the pTRMs induced in a sample by a known laboratory field with the components of the natural remanence that are lost by heating through the same temperature ranges. It provides a consistency check by giving estimates of the original field strength from the remanence induced in each of several temperature ranges. The high blocking temperature components are generally the most stable and so give the most reliable results.

The idea is that, if a sample has a simple thermoremanence and is chemically unaffected by heat, then the Thellier method gives an estimate of the strength of the original field in which it cooled; the error arising from the dependence of blocking temperature on cooling rate is assumed to be small (see Problem 25.2). However, there are some doubts and difficulties. Chemical changes during laboratory heating can usually be recognized, but there is a possibility that the natural remanence of an apparently unaltered igneous rock is not a simple thermoremanence but, at least partly, a chemical remanence, induced by the field during development or modification of the magnetic minerals. In these cases, the pTRM observations generally indicate that something is wrong by disagreement between the field estimates from different pTRM components. Unfortunately rather few rocks give ideal results; ancillary tests are applied to determine whether field estimates from some components are nevertheless acceptable.

Several modifications of the Thellier method are now preferred, in most cases designed to reduce the labour-intensive procedure of repeatedly heating, cooling and re-measuring samples. The use of field-free space is now standard, although not used in the original work. Shaw (1974) supplemented thermal measurements with alternating field measurements. Anhysteretic remanence (ARM) is, in some interesting ways, analogous to thermoremanence. ARM is induced by a small, steady field applied while a superimposed alternating field is reduced to zero from a high value, a process that would demagnetize a sample in the absence of the steady field. Partial ARM (pARM) is observed when the steady field is applied only for a limited range of the diminishing alternating field. pARMs are additive in the same way as pTRMs and the 'spectrum' of ARM demagnetization is often very similar to that of TRM. Shaw (1974) showed that comparison of the alternating field demagnetization curves of a sample before and after heating allows identification of parts of the curve that are unaffected by heat and are assumed to give reliable paleointensity estimates.

An assumption intrinsic to the Thellier method is that the different pTRM components are independent. Dunlop and Özdemir (1997) drew attention to the fact that the additivity of pTRMs is essentially a single domain phenomenon, with independent fine grains having individual blocking temperatures. In large grains the individual domains may be blocked at particular temperatures, but they are not independent, so that true multidomain TRM is not additive. The most stable remanence, which is of interest to paleomagnetism, is of the pseudo-single domain type (Section 25.2) and the assumption is that this follows Thellier's additivity principle well enough to allow paleointensity measurements using pTRM measurements.

Relative field intensities can be inferred from sediment cores, as in Fig. 25.7, either by assuming that the sediment is uniform over the depth range of interest or by using susceptibility as a measure of the abundances of magnetic minerals. Absolute intensity measurements by the Thellier or Shaw methods necessarily rely on igneous rocks with simple histories and suitably small magnetic grains. These are increasingly difficult to find as one looks further back in time.

It is the oldest rocks that are of greatest interest, as indicators of the very early field. Available data for the age range 2.0 to 3.5 billion years (Hale, 1987; Halls *et al.*, 2004; Macouin *et al.*, 2004; McArdle *et al.*, 2004) suggest that the field was then systematically weaker than in more recent times, as might be expected from the discussion of dynamo power in Section 22.7. It is more or less impossible to be sure of paleointensity results from such old rocks and the field varies continuously anyway, so an enormous amount of data would be needed for a secure conclusion. But there is one important conclusion that is secure: there has been a field comparable to the present one continuously for at least 3.5 billion years.

25.6 Polar wander and continental drift

Direct observations of the Earth's rotational axis by conventional astronomical methods, and now more accurately by satellite observations and very long baseline interferometry (VLBI), show that there is a drift of the North Pole towards 79 °W at about 11 cm/year. Strictly, it is the Earth's features that are moving relative to the rotational axis, which is unaffected by this process. The motion is superimposed on the 14-month and 12-month wobbles that have a total amplitude of about 5 m (Section 7.3). The polar drift is attributed to an axial readjustment by the mass redistribution of post-glacial rebound (Section 9.5), a consequence of asymmetrical glaciation of polar regions during the last ice age. It is a transient phenomenon when viewed on the time scale of the tectonic processes that cause polar wander and continental drift, and causes a total movement of the pole of no more than a few kilometres. However, the rebound drift obscures from direct observation by modern geodetic methods the longer-term polar migration that is studied by paleomagnetism. Satellite and VLBI methods are used to measure tectonic movements of the plates relative to one another, but cannot distinguish the absolute tectonic motion from the rebound effect.

On the other hand, paleomagnetism measures continental movements, relative to the pole, over hundreds and even thousands of millions of years. The paleolatitude of a rock at the time of its formation is given by the dip angle of its magnetization, using Eq. (24.11), and its orientation is indicated by the direction of the horizontal component of magnetization. Thus, the angular distance and direction to the pole are determined and its position, relative to the sampled rock, can be plotted on a globe, or on a projection of one. Sufficient observations are made to average out the secular variation and, assuming validity of the axial dipole principle (while acknowledging the small far-sided effect discussed in Section 25.3), the paleomagnetic pole determined in this way is also the geographic pole. Measurements on a series of samples of different ages from the same land mass give a series of pole positions that mark a path, called an apparent polar wander path. Thus, a primary observation of paleomagnetism is polar wander, which can be observed unambiguously for any land mass that remains coherent. A typical rate of polar wander is 0.3 ° per million years and, for rocks younger than about 20 million years, this cannot be separated satisfactorily from the scatter due to secular variation (Fig. 25.4). Deviations from the present pole are apparent at about 30 million years and become clearer with increasing age.

Greatest interest in polar wander curves arises from comparisons of pole paths for different continents, as in Fig. 25.10. Since, by the axial dipole principle, the averaged pole was in the same position at any particular time for all continents, the differences between pole paths indicate relative drift of the continents. But, whereas polar wander, relative to a particular land mass, is completely specified by paleomagnetic observations, continental drift is subject to an ambiguity in longitude. Given the position of a pole for a continent, we can place the continent on a globe in latitude and orientation, but its longitude is arbitrary. Thus, the longitude difference between two continents cannot be deduced solely from paleomagnetic observations. For continental movements over the last 150 million years we have vital additional information in the magnetic stripes on the ocean floors (Section 12.3) and the relative movements of continents are fully resolved. For earlier periods, from which there is no surviving ocean floor, the longitude ambiguity remains a formal problem, and less precise methods are used to resolve it. The record is less secure for the Precambrian period, more than 500 million years ago, but continental reconstructions back to 2.5 billion years have been presented (Pesonen *et al.*, 2003).

The first two pole paths to be compared were for Europe and North America (Fig. 25.10). As was immediately recognized, the simplest explanation for their divergence is that the two continents were once adjacent and drifted apart by the opening up of the Atlantic Ocean basin. This is an interesting example of the longitude ambiguity problem mentioned above. The curves are

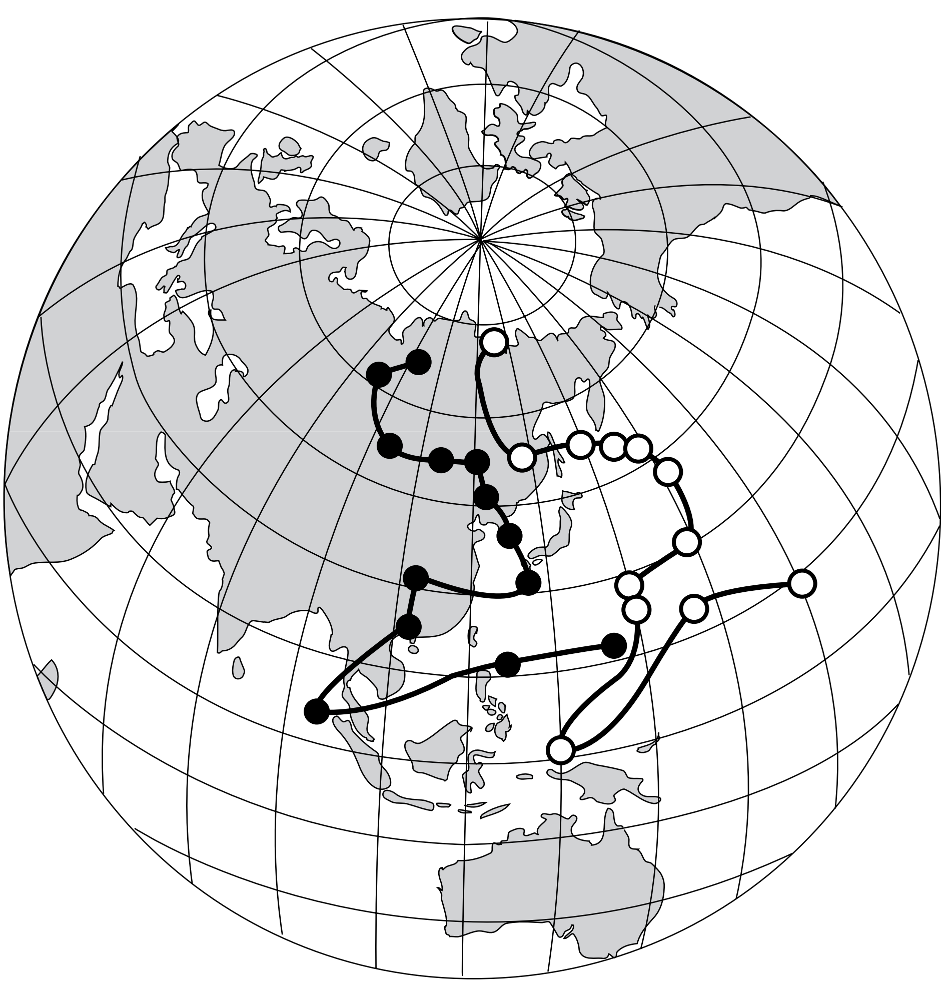


FIGURE 25.10 Pole paths for Europe (open circles) and North America (solid circles) for a 250 million year period from the Jurassic to the Ordovician. These pole paths can be made to coincide by closing the Atlantic Ocean and joining the continents as in Fig. 25.11. Figure redrawn from Van der Voo (1990).

explained by a relative longitudinal drift of the two continental blocks, but paleomagnetism cannot distinguish this from rotation. However, rotation sufficient to explain the separated pole paths would give overlap of the continents and must be discounted. In any case we have marine magnetic anomalies to demonstrate that the opening of the Atlantic basin is the correct interpretation.

The displaced pole paths for Europe and North America revived the idea of continental drift. It was slow to gain general acceptance, but the subsequent demonstration of the greater movement of Australia relative to the northern continents removed much of the lingering doubt. The idea that the continents bordering the Atlantic were once juxtaposed and rifted apart has a long history. Similarities of continental shapes and the rocks on corresponding margins provided the early drifters with the evidence they used to argue for continental mobility. The early discussions focussed mainly on the work between about 1910 and 1930 of Alfred Wegener, a German meteorologist. Fig. 25.11 is a more recent fit of the continents bordering the Atlantic, joined in the manner that Wegener favoured. Another important pioneer was the South African geologist A.L. DuToit, whose 1930s reconstruction of Gondwanaland, the composite southern continent comprising

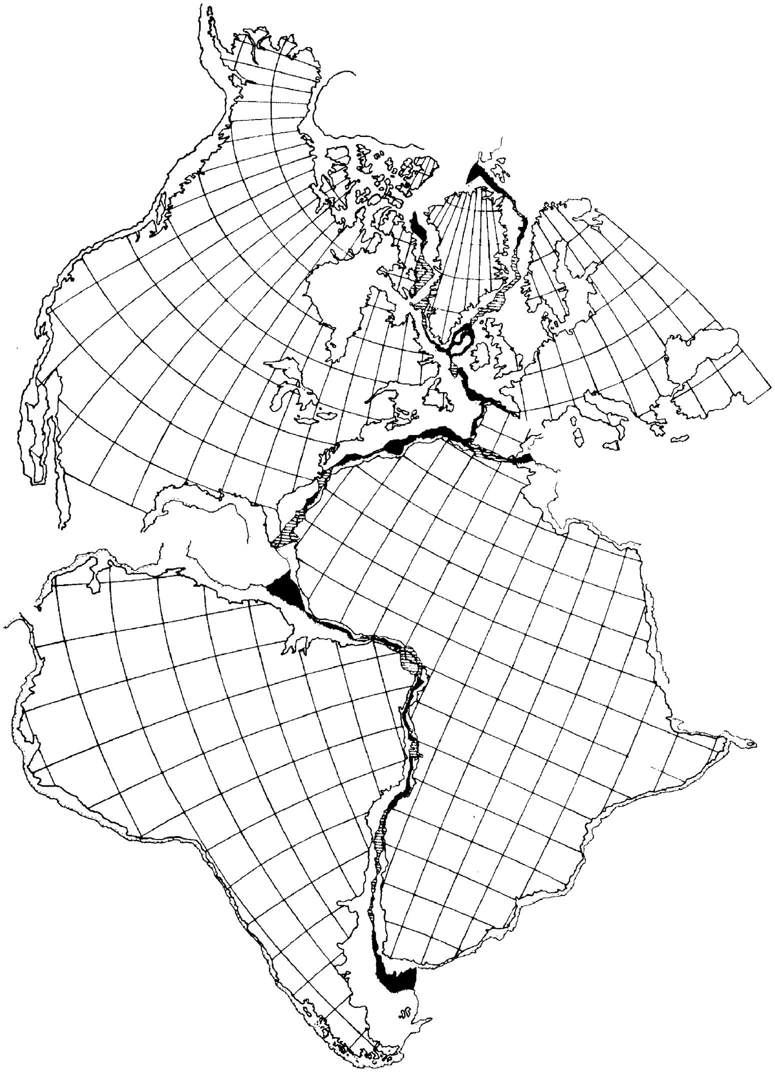


FIGURE 25.11 A fit of the 500 fathom (914 m) depth contour for continents around the Atlantic Ocean. Shaded areas represent overlap and black areas are gaps. Redrawn after Bullard *et al.* (1965).

South America, Africa, India and Australia in a cluster around Antarctica, has been fully vindicated by paleomagnetism (Fig. 25.12).

In early paleomagnetic work, when the subject was still contentious, it was considered important to demonstrate agreement between magnetically measured paleo-latitudes and climate. Glaciation is a particularly effective climatic indicator. Although the global climate is variable, as ice ages demonstrate, it is consistent at any one time. Thus, the ice age glaciations of North America and Eurasia, as well as the southern Andes and South Island, New Zealand, were synchronous. They do not indicate very rapid polar wander but simultaneous cooling. Heavy glaciation occurred in the late Carboniferous and Permian periods, and Holmes (1965) showed that the evidence for this in the southern continents is consistent with their clustering about Antarctica at that time (Fig. 25.13).

Plate tectonics (Chapter 12) is now central to global Earth science, with a wide range of observations to elucidate the details. But the starting point was the paleomagnetic demonstration of continental drift and everything else followed. The apparent relegation of the

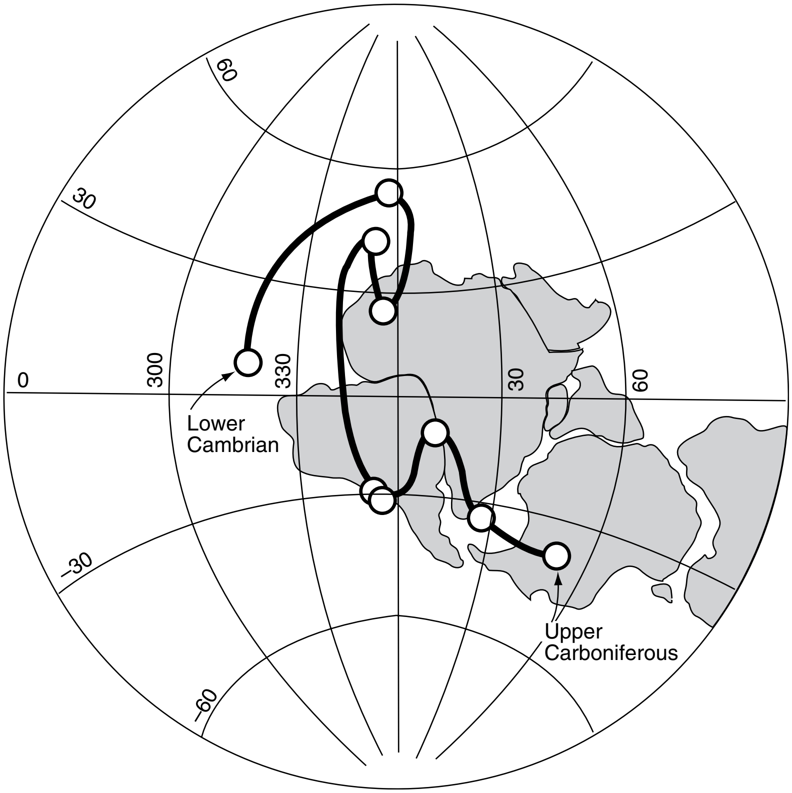


FIGURE 25.12 The continents of Gondwanaland reassembled around Antarctica, to give a common pole path for South America, Africa, India and Australia for Carboniferous to Cambrian periods. Redrawn part figure from Van der Voo (1992).

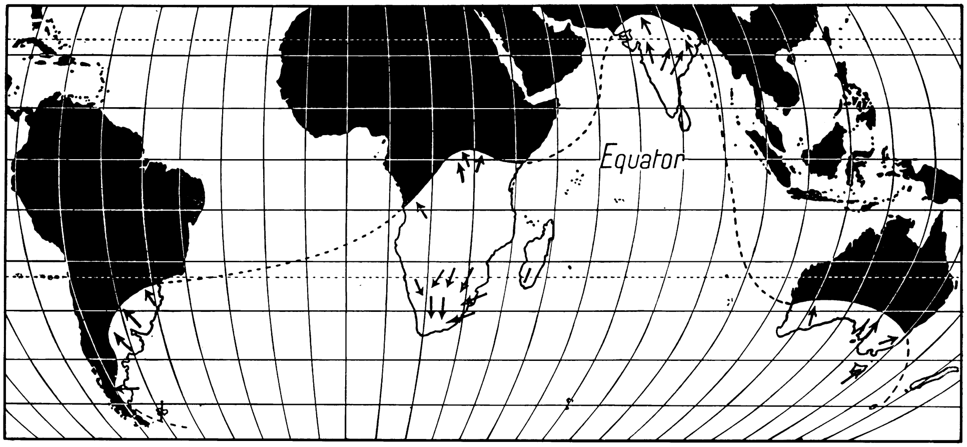


FIGURE 25.13 Glaciation of the southern continents during the Late Carboniferous period. Arrows indicate directions of ice movement. The glaciation and directions of ice movement are consistent with the clustering of these land masses around Antarctica as a single super-continent, Gondwanaland, as in Fig. 25.12. Reproduced, by permission, from Holmes (1965).

drift almost to the last section of this text is not an indication of a minor role but recognition that it is background information that passes as common knowledge. Over no more than a decade or two, paleomagnetism was transformed from an exciting new sub-discipline, overturning the standard view of the Earth, to a basic tool.

26

'Alternative' energy sources and natural climate variations: some geophysical background

26.1 Preamble

Solid Earth geophysics offers assessments of energy sources alternative to fossil fuels. We gain some insight on the accessibility and geophysical effects of exploiting any particular energy source by comparing the contemplated scale of its use with the corresponding natural dissipation. The analyses and discussions of earlier chapters are used to address these issues. We consider also the astronomically induced climate variations that must be distinguished from the consequences of fossil fuel use. This is background material to the environmental questions that are primarily problems for atmospheric science and to the resource question of fossil fuel exhaustion that is a central problem for exploration geophysicists. Energy sources considered are solar, wind, tidal, ocean wave, hydroelectric and geothermal, but not nuclear. We neglect also the possibilities of biofuel, merely noting that if it is to make a major contribution it will require a large fraction of the land area to be devoted to appropriate crops.

The energy dissipations by various natural processes (Section 26.2), give an indication of their availability for exploitation, and especially the magnitudes of their potential contributions. The comparison with the human use of energy draws attention to the fact that human activity is a global geophysical scale phenomenon. This means that major adjustments to it cannot be made easily in a controlled way. But major adjustments to energy use are inevitable and if they are uncontrolled they are likely to be painful. We examine here some of the basic facts that are needed for informed judgements of the problems and potential solutions. Our discussion is restricted to fundamental questions of energy availability in principle and does not address technical problems of exploitation.

Energy production and use are well documented because most of the production of fuels and electricity occurs in industrial scale operations that are monitored by government authorities. The components of global energy use, itemized in Tables 26.1 and 26.2, are tabulated by the United Nations Statistical Division, the US Department of Energy and others. The original data appear in several forms and in some cases there are significant discrepancies. Generally fuels are assessed in petajoules (1 petajoule/year $= 3.1688 \times 10^7$ W). In American usage an alternative is quads (1 quad $= 10^{15}$ British thermal units $= 1055$ petajoules). Notional equivalents are 1 barrel (159 litres) of oil $= 6.1 \times 10^9$ J and 1 tonne (10^3 kg) of 'standard' coal $\approx 4.3 \times 10^{10}$ J (variable). Electricity generation is quoted in millions or billions of kilowatt-hours/year (10^6 kW-h/year $= 1.14 \times 10^5$ W). All values are here converted to average use in terawatts (10^{12} W) for

Table 26.1 Average global use, in year 2005, of primary energy sources (terawatts) and annual percentage increase

Solids (coal, lignite, peat)	3.98	4.5%
Oil, liquefied gas	5.21	1.9%
Gas	3.42	2.0%
Primary electricity (see Table 26.2)	0.69	2.3%
'Traditional' fuels[a]	~1.1	4.0%
Total	14.40	2.8%

[a] wood, animal waste, etc.

Table 26.2 Global electricity generation (terawatts), year 2005, and annual percentage increase

Primary:	hydroelectricity	0.34	4.0%
	nuclear	0.32	0
	other[a]	0.034	7.3%
Secondary:	thermal[b]	1.24	3.3%
Total		1.93	2.9%

[a] geothermal, wind, solar, tidal, wave generation
[b] coal, oil and gas-fired generation

convenience of comparison. Care is required in some cases, such as the production of electricity by means other than the combustion of fossil fuels. Some tabulations include multiplying factors to give estimates of the fuel that would be required for the same generation rather than electric power itself. That is not done here. Pumped storage generation is also excluded because that is produced from energy otherwise accounted for.

The significance of possible alternative energy sources must be judged against the energy demand in Tables 26.1 and 26.2. The least certain of these numbers is the 'traditional' fuels entry in Table 26.1, which is mainly wood and animal wastes collected for heating and cooking. Their use is still increasing by about 4% per year. The total energy demand is increasing at about the same rate and is probably limited only by a combination of availability and affordability. Availability poses scientific questions to which we draw attention; any active control of demand requires political and social decisions.

Scientific input is essential also to the other problem discussed in this chapter: natural climate variations. We make only fleeting reference to the greenhouse gas problem, but discuss the whole Earth/astronomical considerations. The Earth's closest approach to the Sun on its elliptical orbit occurs on 4 January, the height of the southern hemisphere summer. At this time 7% more sunlight falls on the Earth than at the height of the northern hemisphere summer. The annual cycle of global average temperature does not fully reflect this difference for several reasons, but especially because the ratios of land to sea areas are quite different over the two hemispheres. The precession of the Earth (Section 7.2) is slowly changing this situation, so that in 10 000 years time the Earth will be closest to the Sun in the northern hemisphere summer (as it was 10 000 years ago). Then the northern hemisphere will experience hotter, shorter summers and colder, longer winters. Even if there are no other changes, this fact alone will have a major impact on the climatic pattern. It is the most obvious of the astronomically driven climate variations, the Milankovitch cycles (Section 26.4).

We have another perspective on astronomical effects by noting that the Earth–Sun distance oscillates with the lunar month. The Earth and Moon are orbiting their common centre of mass and it is this centre of mass that follows an elliptical path about the Sun. The motion of the Earth about this path is slight compared with the annual oscillation of the Earth–Sun distance caused by the ellipticity of the orbit, but its period cannot be confused with any other. The fact that a lunar period has been detected in satellite observations of tropospheric temperature (see Section 26.4) provides unambiguous evidence that this monthly oscillation has a climatic effect. The atmospheric response is complicated and not a simple radiative balance, but involves changes in atmospheric circulation. It presents a challenge to atmosphere–climate modelling.

26.2 Natural energy dissipations

That the human use of energy (Table 26.1) is of geophysically significant magnitude is seen by comparing it with the natural dissipations in Table 26.3. Some of the quantities in this table are well measured; others are guesstimates based on assumptions that merit scrutiny and are discussed below, but they are accurate enough for the essential conclusion to be clear. Only the solar radiative power completely dwarfs our energy use and wind is a clear second. Harnessing a few terawatts of solar energy could have no significant impact on it, but the same is not true for the other entries in the table, with the probable exception of wind energy. Inclusion in the table is not an implication that any particular form of energy is exploitable, even in principle; accessibility and consequences of use are considered in the next section.

The deep global processes in Table 26.3 are subjects of Chapters 20, 22 and 24 and tidal friction is discussed in Chapter 8. The listed solar power is the product of the solar constant and the cross-sectional area of the Earth, with the fraction reaching the surface 0.56 of the total at the top of the atmosphere. This leaves four entries in the table that require some explanation, especially as they are the least certain.

Table 26.3 Natural energy dissipations (terawatts)

Deep global processes:		
Geothermal flux	global	44.2
	land areas	9.6
Tectonics		8.0
Geodynamo		~0.3
Surface/atmospheric processes:		
Solar power	top of atmosphere	1.75×10^5
	surface	9.8×10^4
	land surface	2.8×10^4
Tides		3.7
Wind	global	434
	land	126
Waves[a]		~5
River flow		6.5
Atmospheric electric current		9×10^{-4}

[a] Assumes rms peak-to-peak wave height of 2 m on 10^8 m of coastline

For the purpose of estimating the energy dissipation by wind we are interested only in friction with the Earth's surface and not internal dissipation by atmospheric turbulence. The ocean wave energy in the table is wind energy transferred via the oceans to shorelines, but we cannot infer that the transfer of wind energy to land areas is similar, because topography has a strong influence. We need an independent measure of the atmosphere–land coupling. An indication of this is obtained by considering the length of day variations caused by the atmospheric circulation (Fig. 7.4). From the close parallelism of the two curves in this figure it is evident that the coupling of the Earth to the atmospheric motion is tight enough to cause changes of order 1 millisecond in the length of the day over a period as short as 15 days. Variations in the atmospheric motion are rapidly communicated to the surface. We can use this observation to estimate the rate of energy dissipation by the frictional contact.

The change in total rotational energy resulting from a transfer of angular momentum between the atmosphere and the solid Earth, moments of inertia C_A and C_E, with angular speeds ω_A and ω_E, subscripted 1 and 2 for initial and final states, is

$$\Delta E = (1/2)C_E(\omega_{E1}^2 - \omega_{E2}^2) + (1/2)C_A(\omega_{A1}^2 - \omega_{A2}^2). \tag{26.1}$$

Applying conservation of angular momentum,

$$C_E(\omega_{E1} - \omega_{E2}) + C_A(\omega_{A1} - \omega_{A2}) = 0, \tag{26.2}$$

we can rewrite ΔE in two alternative ways that are useful to the present discussion:

$$\begin{aligned}\Delta E = {} & (1/2)C_E(1 + C_E/C_A)(\omega_{E1} - \omega_{E2})^2 \\ & - C_E(\omega_{E1} - \omega_{E2})(\omega_{A2} - \omega_{E2}),\end{aligned} \tag{26.3}$$

$$\begin{aligned}\Delta E = {} & (1/2)C_E C_A/(C_E + C_A) \\ & \times \left[(\omega_{A1} - \omega_{E1})^2 - (\omega_{A2} - \omega_{E2})^2\right].\end{aligned} \tag{26.4}$$

The first term of Eq. (26.3) gives the energy dissipation if the final state is one of precise corotation of the Earth and atmosphere, that is, $\omega_{A2} = \omega_{E2}$. Then, with $\omega_E = 7.292 \times 10^{-5}$ rad s^{-1}, $C_E = 8.036 \times 10^{37}$ kg m^2, $C_A = 1.38 \times 10^{32}$ kgm^2 (Table A.4, Appendix E) and $|\omega_{E1} - \omega_{E2}|/\ \omega_E = 1$ millisecond/24 hours $= 1.16 \times 10^{-8}$, we have $\Delta E = 1.67 \times 10^{19}$ J. Noting that the phase lag between the curves of Fig. 7.4 is no more than 15 days (1.3×10^6 s), we assume that the rate of energy dissipation is at least 1.67×10^{19} J/1.3×10^6 s $= 12.9 \times 10^{12}$ W. The second term in Eq. (26.3) may have either sign. The reason is clearer in Eq. (26.4), which shows that the energy change is simply proportional to the change in the square of the zonal wind speed. However, its magnitude is more readily calculated from Eq. (26.3), using the length of day observations. The speed of a zonal wind of strength sufficient to dissipate the 13 terawatts estimated from Eq. (26.3) is

$$\nu \approx |\Delta\omega_A - \Delta\omega_E|(\pi/4)R_{\text{Earth}} = \Delta\omega_E(1 + C_E/C_A) \times (\pi/4)R_{\text{Earth}} = 2.5 \text{ m s}^{-1}, \quad (26.5)$$

where $(\pi/4)$ is a factor to allow for the latitude variation.

Equation (26.5) gives the zonal wind fluctuation, averaged over the whole atmosphere, that is required to cause the length of day fluctuations in Fig. 7.4. The corresponding energy dissipation, calculated above, is 12.9 terawatts. This is the measure of the atmosphere–earth coupling that we need: a global average wind speed of 2.5 m s^{-1} dissipates about 13 terawatts by its interaction with the Earth. We use this as a calibration of the energy dissipation by interaction of the overall global wind pattern with the body of the Earth. The 2.5 m s^{-1} estimate is not the surface wind speed but the average zonal motion of the atmosphere as a whole. Noting that the energy varies as the square of wind speed in Eq. (26.4), we need only a value of the rms wind speed of the whole atmosphere to estimate the energy dissipation that wind turbines could, in principle, intercept. It is obviously much more than 2.5 m s^{-1} and we adopt, from Archer and Jacobson (2005), 14.3 m s^{-1} as an approximate value. On this basis the dissipation is $(14.3/2.5)^2 \times 12.9$ terawatts $= 434$ terawatts. This is the global value in Table 26.3. It is not immediately clear how this total is divided between land and sea areas. The land is rougher, but wind speed over it is generally lower. The land value in Table 26.3 assumes a fraction proportional to its area. It exceeds the human use of energy by a factor of nearly 9.

We turn now to wave energy. Waves of peak-to-peak amplitude a propagating in deep ocean have energy per unit area

$$E/A = a^2\rho g/4, \quad (26.6)$$

where $\rho = 1025$ kg m^{-3} is sea water density and $g = 9.8$ m s^{-2} is gravity. Waves of frequency f travel in deep water with phase speed $v = g/2\pi f$ (obtained by putting $kh \gg 1$ in Eq. (14.53) and substituting $k = 2\pi f/v$), but the wave energy travels with the group speed $u = v/2$,

$$u = g/4\pi f. \quad (26.7)$$

For a dominant wave period $1/f = 10$s, which is common for ocean swells, $u = 7.8$ m s^{-1}. Thus the rate at which wave energy reaches unit length of shore line parallel to the wave fronts is

$$(E/A)u = a^2\rho g^2/16\pi f. \quad (26.8)$$

Assuming that reflected energy is negligible, but that waves in the deep ocean are randomly oriented with respect to the nearest coastline (introducing a factor $2/\pi$), and that the total global coastline exposed to open ocean waves is $L = 10^8$ m (2.5 great circles), the global total power dissipated by waves of rms peak-to-peak amplitude 2 m is

$$\text{Wave dissipation} = (E/A)(2u/\pi)L = 5 \times 10^{12} \text{ W}, \quad (26.9)$$

as in Table 26.3. This appears to be a generous estimate.

The equations for wave energy and speed used here apply to open ocean situations, that is to water depth much greater than wavelength. As waves approach shallow water and slow up, they are refracted so that the wave fronts are more nearly parallel to the bottom contours and adjacent coastline. However, this complication does not affect the calculation because we have calculated the flux of wave energy from deep ocean into shallow water, where it all turns on to the coastline. We have not considered the dissipation of energy in the open ocean, regarding it as inaccessible, and conclude that the wave energy reaching coastlines cannot become a major contributor to usable power.

To the extent that it is available, hydroelectricity is an ideal power source. The energy is stored indefinitely in a convenient form and its use can be continuously adjusted. The global total rainfall on land is about $10^{14}\,\mathrm{m}^3$/year and the fraction that flows to the sea in rivers and streams is about 25%, giving a mass flow $\mathrm{d}m/\mathrm{d}t = 2.5 \times 10^{16}$ kg/year $= 7.9 \times 10^8\,\mathrm{kg\,s}^{-1}$. If we make the simple assumption that this flows from an average elevation $h = 840$ m, the mean height of all land, then the gravitational energy release is

$$\text{River power} = (\mathrm{d}m/\mathrm{d}t)gh = 6.5 \times 10^{12}\ \mathrm{W}. \quad (26.10)$$

This is probably an underestimate because rainfall is generally greater at higher elevations. But the striking conclusion is that the current hydroelectric power generation (Table 26.2) is already 5% of this hypothetical limit. The limit itself is less than half of the current energy use.

Thunderstorms maintain an electric charge on the Earth that continuously leaks to the ionosphere, with a global current of about 2000 A. The potential difference between the ground and the ionosphere is about 450 000 V, so that the ohmic dissipation by this fine weather current is 9×10^8 W. This is only a very small fraction of the energy of the storms that maintain it and is of no interest here.

Rough as some of these numbers are, energy dissipation by wind must be seen as a clear second to solar energy, with the other contenders more than an order of magnitude smaller.

26.3 'Alternative' energy sources: possibilities and consequences

The variation of solar radiation with the sunspot cycle is about 0.15% (see Section 26.4). Although this effect appears to be slight, and evidence of an 11-year cycle in climate is unconvincing, the variation in radiation received by the Earth on account of solar variability is 20 times the 14 tW of human energy use. There are several implications; an important one is that the thermal effect of energy use is inconsequential when considered in a global perspective. Concerns about global warming have nothing to do with heat release but only with changes to the infra-red opacity of the atmosphere. It is also clear that no conceivable harnessing of solar energy would have a direct climatic implication. We examine the other entries in Table 26.3 in the light of the principle that natural dissipations give a measure of the availability of 'alternative' energies. We consider also the consequences of use.

Power generation from tides is very limited, although a tidal power station has operated in the Rance estuary at St. Malo, on the Atlantic coast of France, for many years. Even at the few sites around the world with comparably large tides, the available head of water is very small by the standards of hydroelectric power generation. The requirement for very large turbines and multiple impoundments of water, or an ancillary pumped storage facility if power is to be maintained through the tidal cycle, make the economics of tidal power doubtful, but such situations can change, so we examine the possibility.

Tidal power has appeared attractive in principle because it has been assumed that the local environmental impact is the only negative consequence and that there is no global effect. However, tidal energy is derived from the Earth's rotation. As discussed in Chapter 8, tidal friction slows the rotation and causes the Moon to recede from the Earth. Natural tidal friction dissipates rotational energy at a rate that is about a quarter of the present human use of energy, so, if a major conversion to tidal power were possible, it would have a first order effect on the rotational slowing. This would still be very gradual (tides are causing the length of day to increase by 2.4 milliseconds per century), and, even with serious harnessing, the time scale for major change would be hundreds of millions of years. There is a remote possibility that accelerated slowing of the rotation would affect the geodynamo (on a time scale of thousands of years), but motions in the core are so rapid compared with any contemplated rotational change that we consider this to be unlikely.

Tidal friction causes a phase lag, δ, of the tide, relative to the position of the Moon or Sun (Fig. 8.4), and dissipation is proportional to $\sin 2\delta$. The satellite-measured phase lag is 2.9°. The maximum possible dissipation would occur for $\delta = 45^\circ$ ($\sin 2\delta = 1$) and if the tidal amplitude were unaffected the dissipation would be $1/\sin 5.8^\circ = 9.9$ times the present rate, giving 36×10^{12} W. But a reduction in tidal amplitude would accompany energy extraction and the theoretical limit is nearer to 20×10^{12} W. Although the energy reservoir is extremely large, it is accessible only in a very limited way. It is not a super-abundant power source, even in principle.

The harnessing of tides would be 'mining' energy in the sense of a permanent and irreversible extraction of energy from a finite, although vast, source. The harnessing of wind, waves and river flow is qualitatively different. All three are by-products of solar energy and all of the energy that is available, in principle, is dissipated naturally anyway. We may intercept and make use of some of it, with local environmental consequences but no effect on the global energy balance.

Wind energy is conveniently exploitable at elevations up to 100 m or so, in an atmospheric boundary layer, within which wind speed increases with height. This layer is responsible for the natural dissipation estimated in Section 26.2 and energy is carried down into it from greater heights. The consequence of extracting wind energy, with turbines in the boundary layer, is subtly different from the effect of obstacles, such as buildings and trees, but the difference is not important. It is obviously impossible to remove all of the energy from the boundary layer because that would completely stop it, leaving no energy to be extracted and causing a new boundary layer to form above it. This confirms two earlier points. The wind energy is generated high in the atmosphere and is dissipated anyway, with or without man-made structures and there is no global environmental consequence to dissipation in man-made structures. The second point applies to all of the energies in Table 26.3: it is possible to 'capture' only a fraction and for a viable source the required fraction needs to be small. In response to the question 'How small?', we note that, in the case of wind, by our estimates, 12% would suffice to satisfy global energy demand if energy extraction is confined to land areas. Archer and Jacobson (2005) made a more direct estimate of available wind energy from records of wind speeds at 80 m elevation, a standard height for wind turbines. They concentrated attention on continental areas of high average wind speeds, v, because extractable energy depends on v^3 (the rate at which air mass passes through a turbine is proportional to v and its kinetic energy per unit mass is proportional to v^2). The essential conclusion of Archer and Jacobson is that sufficient energy is indeed accessible, with existing technology, for wind to become the dominant world energy source.

Ocean waves represent a more concentrated form of energy than the wind that drives them, but we are interested only in the dissipation that occurs at coastlines. The calculation leading to Eq. (26.9) gives an energy flux of 50 kW per metre of wavefront for waves of 2 m peak-to-peak amplitude and 10 s period. This concentration of mechanical energy makes it appear attractive from the perspective of small-scale engineering, but on a global scale the 5 tW entry in Table 26.3 makes it evident that, irrespective of technical problems, waves have no prospect of becoming a major player in the energy game.

As pointed out in the previous section, if every drop of water that flows to the sea in rivers and streams were to flow all the way through turbines of 100% efficiency, the total power generation (Table 26.3) would be less than half of the present total energy use and only 20 times the current hydroelectric power generation (which continues to increase). This is a comment on the ready accessibility of river power, but it is also an indication that the most favourable sites are already in use. The total capacity of the world's hydroelectric dams is about 7×10^{11} m^3 and this water storage on land has lowered sea level by about 2 mm (perhaps 10% of the effect of all the smaller dams). This is inconsequential. The increased moment of inertia of the Earth is far below the level that would cause an observable effect on rotation. The environmental consequences of harnessing river power are local, not global. But, although the energy of river flow is very accessible, in an engineering sense, the total in Table 26.3 makes it clear that this cannot, in principle, become a dominant energy source.

Geothermal power stations in Iceland, Italy, New Zealand and California use steam from ground water in volcanic areas. Although they are a valuable source of power in these areas, they require very special geological conditions. Interest in the wider use of deep heat assumes that it is possible to extract heat from hot, dry rock at depth in the crust. In geologically stable continental regions the ambient heat flux averages about 0.065 $\mathrm{W\,m^{-2}}$ and the temperature gradient is typically 25 K/km, so that prohibitively deep drilling would be required to reach anything more than very low-grade heat. Only areas with unusually high temperature gradients offer any prospect of economic heat extraction and, as in Fig. 20.4, these are restricted to geologically young igneous regions. We can take a closer look at this problem by modifying Eqs. (20.15) and (20.16) to model an area with a uniform crust, having a layer of granite extending from the surface to depth z_0. We assume heat generation $\dot{q} = 2.8 \times 10^{-6}\ \mathrm{W\,m^{-3}}$ (corresponding to $1.05 \times 10^{-9}\ \mathrm{W\,kg^{-1}}$, as in Table 21.3), conductivity $\kappa = 2.5\,\mathrm{W\,m^{-1}\,K^{-1}}$ and a heat flux $\dot{Q}_{\mathrm{MC}} = 0.025\ \mathrm{W\,m^{-2}}$ from the mantle. Relative to the surface value T_0, the temperature at depth z is

$$T(z) - T_0 = \dot{Q}_{\mathrm{MC}} z/\kappa + (\dot{q}/\kappa)(z_0 z - z^2/2). \quad (26.11)$$

At a suitable drilling depth of 3 km this gives

$$T(3\ \mathrm{km}) - T_0 = 25\mathrm{K} + 3.36\mathrm{K} \times z_0(\mathrm{km}), \quad (26.12)$$

so that 20 km of granite would give an equilibrium temperature only 90 K above that at the surface (or 150 K at 5 km depth). We see that rock in thermal equilibrium at accessible depths could be usefully hot only if it is far more radioactive than normal granite or if it extends to implausible depths.

The conclusion from these equations is that hot dry rock geothermal power projects are possible only in areas of high heat flux arising from the residual heat of igneous activity. The heat generation in the rock itself is only marginally relevant and we can note that, without loss of heat, the temperature rise in granite due to its own radioactivity would be only about 40 K per million years. Moreover, a prodigious thickness would be required to stop it diffusing out. But the geological requirement is not as restrictive as that for conventional geothermal power generation, which taps ground water superheated by more recent (or current) volcanic activity.

We can also look at the energy balance of a hot dry rock project. If we suppose that circulation of water between boreholes in fractured rock extracts sufficient heat to cool 1 $\mathrm{km^3}$ of rock by 100 K, before the heat degrades too far to be effective, and that the average thermodynamic efficiency of power generation over the usable temperature range is 20%, then it would produce 100 megawatts for 16 years. That volume of rock would then be thermally exhausted. On a time scale of 10^4 to 10^5 years, partial recovery by diffusion of heat from hot surroundings, including deeper rock, is possible but not assured. Like the conventional 'hot wet rock' geothermal power generation, this would be mining heat of igneous origin, not exploitation of the geothermal flux listed in Table 26.3 and discussed in Section 20.3. But the heat is abundant in restricted areas and the fundamental limitation is its distribution. Hot dry rock could provide geothermal power in geologically youthful igneous provinces, although these are sufficiently limited to keep it out of the big league in the energy game.

By a process of elimination we reach the conclusion that, disallowing nuclear power, the only replacement sources that can, in principle, supply energy on the scale of the current use of fossil fuels are solar and wind. That they are also the most widely distributed sources is not incidental; we are not considering concentrated sources and wide distribution is necessary simply to meet the total requirement. It follows that, as 'alternative' sources come into play, power generation will become increasingly localized in smaller scale operations than has been conventional. Barring some unanticipated discovery, these two sources must, between them, eventually take over. The perceived difficulty of discontinuous or erratic availability is a technical problem, to which solutions are at hand, including the already well established pumped storage system.

26.4 Orbital modulation of insolation and solar variability

Precise satellite observations of the luminosity of the Sun now extend over more than two 11-year cycles of sunspot numbers. Willson and Hudson (1991) appear to have been first to demonstrate that it varies by about 0.15%, with maxima coinciding with sunspot maxima. Although the record is short, we have independent evidence that sunspots correlate with, and can be used as an indicator of, solar output. The Maunder minimum in activity (~1645–1715), when sunspots were virtually absent for 70 years, coincided with the 'little ice age' of lowered global temperature, and we note a report by Zhang *et al.* (1994) of correlated luminosity and magnetic activity of ten solar type stars. Although clear reports of an 11-year cycle in global temperature are lacking, variability of the Sun must be allowed as an important contributor to climate change, perhaps the most important one, although not understood. The enhanced greenhouse effect of carbon dioxide and other gases must be seen against a background of such natural effects.

As mentioned in the preamble, the Earth's orbit about the Sun is eccentric ($e = 0.016\,73$), giving a ratio of perihelion (closest) to aphelion (most remote) distances of 0.967 09. The intensity of radiation received by the Earth varies inversely as the square of this factor, being stronger on January 4 than six months later by a factor 1.069. If we make the simple assumption that the Earth behaves as a radiative black body, emitting energy proportional to T^4, the variation in insolation would imply a global average temperature oscillation of 4.6 K. An effect as big as this would be very obvious, but, apart from atmospheric feedback mechanisms, it is mitigated and obscured by the asymmetry of continents and oceans, with a much greater oceanic area in the south. The consequences include greater cloudiness there, but this state of affairs is not permanent. The Earth's axis of rotation precesses (Section 7.2) and the alignment of the hemispheres with maximum insolation interchanges. This is one aspect of the Milankovitch cycles, discussed below.

Surprisingly, one of the weakest of the astronomical cycles has an observed effect. This is the lunar modulation of the tropospheric temperature (roughly the lowest 6 km of the atmosphere) as seen by satellite measurements of thermally excited microwave radiation from molecular oxygen. The Earth and Moon orbit the Sun together and it is the centre of mass of the combination that follows the elliptical orbit about the Sun. The shared centre of mass is a point within the Earth 4671 km from the centre (assuming the Moon to be at its mean distance, 3.844×10^5 km). Thus, there is a monthly oscillation in the Earth–Sun distance amounting to 2×4671 km, a variation of 6.2 parts in 10^5, causing a monthly oscillation in insolation of 1.24×10^{-4}. There is also a contribution by moonlight, which is strongest at full Moon, when the Earth is closest to the Sun, and so reinforces the variation. For this purpose we can assume the albedo (reflectivity) of the Moon to be unity, because absorbed radiation is re-radiated in the infra-red and intercepted by the Earth. (Most of this infra-red radiation is emitted from the hotter, sunlit side of the Moon.) Allowing for the slightly smaller angular size of the Moon than the Sun, as seen from the Earth, the moonlight peak is 2.0×10^{-5} of the solar radiation. This makes the total monthly variation in insolation 1.44×10^{-4}.

A monthly cycle of tropospheric temperature was identified by Balling and Cerveny (1995) in records of satellite-monitored radiation from molecular oxygen. The global average tropospheric temperature is 269 K and oscillates by about 0.01% of this, but does not have a simple phase relationship with the cycle of insolation. Over the lunar month, the temperature peak precedes the maximum in insolation (at full Moon) and there is a complicated latitude variation. At the equator, little effect is seen. At high and low latitudes, in both hemispheres, the temperature variation follows the global trend, but at mid-latitudes (roughly 40° to 60°), the oscillation is reversed. The pattern of atmospheric circulation is varying slightly in a systematic way over the lunar month. Although the possibility of a tidal influence cannot be securely discounted, the obvious driver is radiative forcing, but there is no theoretical explanation. It is useful to note also that Gordon (1994) used the same microwave observations to find a weekly cycle in tropospheric temperature, amounting to 0.01 K (warmer on weekdays) in the northern hemisphere, but not in the south.

The demonstrated sensitivity of tropospheric temperature to very small variations in insolation appears to conflict with the limited evidence for a climatic response to ellipticity of the Earth's orbit, illustrating the complexity of the climate system. Atmospheric feedback mechanisms, compounded with the ocean/continent asymmetry of the hemispheres, make it difficult to isolate cause–effect relationships. This is a reason for interest in the satellite microwave observations. The lunar period is unique. There is no way of attributing the appearance of this period in tropospheric temperature to anything but the Earth–Moon orbital motion. Similarly, the weekly cycle, distinguishing weekdays from weekends, has no astronomical or other natural basis and is unambiguously identified with human activity. The observations provide vital evidence that the atmosphere responds in a comprehensible way to both variations in insolation and the impositions of human activity. We can keep these situations in mind when trying to understand more complicated effects with multiple causes.

Natural climate changes occur on all time scales but variations with characteristic times of a few tens of thousands of years have attracted particular attention. This is the time scale of the major advances and recessions of glaciers and ice-sheets over the last million years or so (the Quaternary period). The idea that ice ages were brought on by diminished insolation arising from orbital changes originated in the 1800s, especially with the work of James Croll, and was given a sound theoretical basis by M. Milankovitch in the early 1940s. With subsequent improvements in dating Quaternary geological events and in calculating details of orbital motion, it has become clear that the astronomical cycles expected by the Milankovitch theory are seen in the observed climatic changes (Berger, 1988; Berger and Loutre, 1992). There is even a report of geological evidence of Milakovitch cycles 500 million years ago (Williams, 1991). However, there are obviously other factors, including changes in the Sun and probably ocean circulation, not all of which are understood.

The precession of the Earth (Section 7.2) causes the axis of rotation to cycle about the normal to the ecliptic plane with a period of 25 700 years. If this were the only variation then in half this time the Earth would be closest to the Sun in the northern hemisphere summer instead of the southern hemisphere summer, as at present. But the orientation of the axis of the orbital ellipse also changes and the combined effect is a climatic precessional period of 21 000 years. It depends on the fact that the orbit is elliptical, but this also changes. The eccentricity varies with a period of about 96 000 years, and when it is very small, climatic evidence of the precession disappears. The other important orbital parameter is obliquity (inclination of the rotational axis to the normal to the orbital plane), which has a 41 000-year period. These effects all interact to produce variations in insolation with latitude as well as season and the Earth responds in a non-linear manner, with varying albedo due to snow and ice cover. Nevertheless, climate indicators, such as oxygen isotope ratios (Section 3.9), give sufficient evidence of Milankovitch periods to convince us that they account for a significant part of the climate variations.

It is difficult to attribute to orbital variations the prolonged extreme glaciations that occurred in the Permian period and also in the late Precambrian period. The possibility of major changes to the greenhouse gas content of the atmosphere must be allowed, but solar variability is also implicated. If we attribute major changes to the Sun itself, we must allow lesser effects to have the same cause and with this much freedom to blame the Sun, our theories become very nebulous.

26.5 A concluding comment regarding 'alternative' energies

Noting the natural energy dissipations in Table 26.3, and acknowledging that, in most cases, only small fractions could be harnessed, we have a very limited short list of potential major contributors of 'renewable' energy. Tide, wave and geothermal generators may be locally useful, but their existence cannot obscure this simple arithmetic. We suggest an important role for hydroelectricity, not because there is any possibility of its coming close to supplying the total energy needs, but because it provides an ideal storage system for energy from the two major intermittent sources, solar radiation and wind. These two are the only candidate renewable resources that can, in principle, provide power on the scale of present use, let alone hoped-for future use.

Appendix A

General reference data

Table A.1 Fundamental physical constants

Speed of light in vacuum	$c = 2.997\,924\,58 \times 10^{8}\,\mathrm{m\,s^{-1}}$
Permeability of free space	$\mu_0 = 4\pi \times 10^{-7}\,\mathrm{H\,m^{-1}}$
Permittivity of free space	$\varepsilon_0 = 1/\mu_0 c^2 = 8.854\,187\,8\ldots \times 10^{-12}\,\mathrm{Fm^{-1}}$
Gravitational constant	$G = 6.6743(7) \times 10^{-11}\,\mathrm{m^3\,kg^{-1}\,s^{-2}}(\mathrm{N\,m^2\,kg^{-2}})$
Planck constant	$h = 6.626\,069\,0(3) \times 10^{-34}\,\mathrm{J\,s}$
Elementary charge	$e = 1.602\,176\,49(4) \times 10^{-19}\,\mathrm{C}$
Electron mass	$m_\mathrm{e} = 9.109\,382\,2(5) \times 10^{-31}\,\mathrm{kg}$
Proton mass	$m_\mathrm{p} = 1.672\,621\,64(8) \times 10^{-27}\,\mathrm{kg}$
Neutron mass	$m_\mathrm{n} = 1.674\,927\,21(8) \times 10^{-27}\,\mathrm{kg}$
Atomic mass constant (^{12}C mass/12)	$u = 1.66\,053\,878(1) \times 10^{-27}\,\mathrm{kg}$
Avogadro's number	$N_\mathrm{A} = 6.022\,141\,8(3) \times 10^{23}\,\mathrm{mol^{-1}} = 6.022\,142\,0(5) \times 10^{26}(\mathrm{kg\,mol})^{-1}$
Gas constant	$R = 8.314\,472(15)\,\mathrm{J^{-1}K^{-1}} = 8.314\,472(15)10^3\,\mathrm{J(kg\,mol)^{-1}K^{-1}}$
Boltzmann's constant (R/N_A)	$k = 1.380\,650(2) \times 10^{-23}\,\mathrm{J\,K^{-1}}$
Stefan–Boltzmann constant ($2\pi^5k^4/15h^3c^2$)	$\sigma = 5.670\,40(4) \times 10^{-8}\,\mathrm{Wm^{-2}\,K^{-4}}$
Faraday constant (e/u)	$F = 9.648\,534\,0(2) \times 10^4\,\mathrm{C\,mol^{-1}} = 9.648\,534\,2(4) \times 10^7\,\mathrm{C\,(kg\,mol)^{-1}}$
Inverse fine structure constant ($2h/\mu_0 ce^2$)	$\alpha^{-1} = 137.035\,999\,68(9)$
Rydberg constant ($m_\mathrm{e}c\alpha^2\,/2h$)	$R_\infty = 1.097\,373\,156\,853(7) \times 10^7\,\mathrm{m^{-1}}$

Table A.2 Unit conversions

1 inch	= 0.0254 m (exact)
1 statute mile	= 1609.344 m (exact)
1 nautical mile	= 1852 m (definition; originally 1 minute of latitude)
1 astronomical unit (AU) (mean Earth–Sun distance)	= 1.495 978 71 × 10^{11} m
1 arc sec	= $\pi/648\,000$ rad = 4.848... × 10^{-6} rad
1 pound (lb)	= 0.453 592 37 kg
1 tonne	= 1000 kg
1 ton (US) (2000 lb)	= 907.184 74 kg
1 ton (Imperial) (2240 lb)	= 1016.0469 kg
1 gallon (US)	= 3.636 77 litres
1 gallon (Imperial)	= 4.545 96 litres
1 sidereal year	= 3.155 815 × 10^{7} s
1 erg	= 10^{-7} J
1 dyne	= 10^{-5} N
1 Gal	= 10^{-2} m s^{-2} [1 mGal = 10^{-5} m s^{-2}]
1 atmosphere	= 101 325 Pa
1 bar	= 10^{5} Pa
1 Poise	= 0.1 Pa s
1 calorie	= 4.1868 J
1 heat flux unit (1 microcalorie/(cm^2 s))	= 4.1868 × 10^{-2} W m^{-2}
1 electron volt (eV)	= 1.602 177 33(49) × 10^{-19} J
1 Gauss	= 10^{-4} T(Telsa) = 10^{5} nT(gamma)
1 Oersted	= $10^{3}/4\pi$ A m^{-1} (Ampere-turn/m)
1 Gauss-cm^3 (magnetic moment)	= 10^{-3}A m^{2}
1 e.m.u. of magnetization	= 10^{3} A m^{-1}

Table A.3 Atomic weights of the naturally occurring elements. Each element is listed with its atomic number, z, its chemical symbol (in parenthesis) and the mean atomic mass, m, in atomic mass units, u (Table A.1)

z	Elements	m	z	Elements	m
1	Hydrogen (H)	1.0079	47	Silver (Ag)	107.868
2	Helium (He)	4.002 60	48	Cadmium (Cd)	112.40
3	Lithium (Li)	6.941	49	Indium (In)	114.82
4	Beryllium (Be)	9.012 18	50	Tin (Sn)	118.69
5	Boron (B)	10.81	51	Antimony (Sb)	121.75
6	Carbon (C)	12.011	52	Tellurium (Te)	127.60
7	Nitrogen (N)	14.0067	53	Iodine (I)	126.9045
8	Oxygen (O)	15.9994	54	Xenon (Xe)	131.30
9	Fluorine (F)	18.998 40	55	Cesium (Cs)	132.9054
10	Neon (Ne)	20.179	56	Barium (Ba)	137.34
11	Sodium (Na)	22.9898	57	Lanthanum (La)	138.9055
12	Magnesium (Mg)	24.305	58	Cerium (Ce)	140.12
13	Aluminium (Al)	26.981 54	59	Praseodymium (Pr)	140.9077
14	Silicon (Si)	28.086	60	Neodymium (Nd)	144.24
15	Phosphorus (P)	30.973 76	61	Promethium (Pm)	
16	Sulphur (S)	32.06	62	Samarium (Sm)	150.4
17	Chlorine (Cl)	35.453	63	Europium (Eu)	151.96
18	Argon (Ar)	39.948	64	Gadolinium (Gd)	157.25
19	Potassium (K)	39.098	65	Terbium (Tb)	158.9524
20	Calcium (Ca)	40.08	66	Dysprosium (Dy)	162.50
21	Scandium (Sc)	44.9559	67	Holmium (Ho)	164.9304
22	Titanium (Ti)	47.90	68	Erbium (Er)	167.26
23	Vanadium (V)	50.9414	69	Thallium (Tm)	168.9342
24	Chromium (Cr)	51.996	70	Ytterbium (Yb)	173.04
25	Manganese (Mn)	54.9380	71	Lutecium (Lu)	174.97
26	Iron (Fe)	55.847	72	Hafnium (Hf)	178.49
27	Cobalt (Co)	58.9332	73	Tantalum (Ta)	180.9479
28	Nickel (Ni)	58.71	74	Tungsten (W)	183.85
29	Copper (Cu)	63.545	75	Rhenium (Re)	186.2
30	Zinc (Zn)	65.38	76	Osmium (Os)	190.2
31	Gallium (Ga)	69.72	77	Iridium (Ir)	192.2
32	Germanium (Ge)	72.59	78	Platinum (Pt)	195.09
33	Arsenic (As)	74.9216	79	Gold (Au)	196.9665
34	Selenium (Se)	78.96	80	Mercury (Hg)	200.61
35	Bromine (Br)	79.904	81	Thallium (Tl)	204.37
36	Krypton (Kr)	83.80	82	Lead (Pb)	207.2 (variable)
37	Rubidium (Rb)	85.468	83	Bismuth (Bi)	208.9804
38	Strontium (Sr)	87.63	84	Polonium (Po)	
39	Yttrium (Y)	88.9059	85	Astatine (At)	
40	Zirconium (Zr)	91.22	86	Radon (Ra)	
41	Niobium (Nb)	92.9064	87	Francium (Fr)	
42	Molybdenum (Mo)	95.94	88	Radium (Ra)	
43	Technetium (Tc)		89	Actinium (Ac)	
44	Ruthenium (Ru)	101.07	90	Thorium (Th)	232.0381
45	Rhodium (Rh)	102.9055	91	Protoactinium (Pa)	
46	Palladium (Pd)	106.4	92	Uranium (U)	238.029

Table A.4 Dimensions and properties of the Earth

Equatorial radius (geoid)	$a = 6378\,136$ m
Polar radius	$c = 6356\,751$ m
Lower mantle radius (PREM)	5701 000 m
Core radius (PREM)	3840 000 m
Inner core radius (PREM)	1221 500 m
Flattening	$f = \dfrac{a-c}{a} = 3.352\,81 \times 10^{-3}$
Volume	$V = 1.083\,20_7 \times 10^{21}$ m^3
Radius of sphere of equal volume	$R_E = 6371\,000$ m
Surface areas:	
Total[a]	$A = 5.100\,655 \times 10^{14}$ m^2
land	1.48×10^{14} m^2
sea	3.62×10^{14} m^2
continents (including margins)	2.0×10^{14} m^2
Gravitational constant × mass (including atmosphere)	$GM_\oplus = 3.986\,004\,415(8) \times 10^{14}$ m^3 s^{-2}
Geoid potential	$W_0 = -6.263\,686 \times 10^7$ m^2 s^{-2}
Mass	$M_\oplus = 5.9723(1) = 10^{24}$ kg
Mean density	$\bar{\rho} = 5515$ kg m^{-3}
Mass of atmosphere	5.28×10^{18} kg
oceans	1.4×10^{21} kg
solid crust	2.8×10^{22} kg
lower mantle	2.94×10^{24} kg
mantle	4.00×10^{24} kg
outer core	1.84×10^{24} kg
inner core	9.8×10^{22} kg
Moments of inertia	
about polar axis	$C = 8.0359(12) \times 10^{37}$ kg m^2
about equatorial axis	$A = 8.0096 \times 10^{37}$ kg m^2
core	$C_c = 0.956 \times 10^{37}$ kg m^2
atmosphere	$C_a = 1.38 \times 10^{32}$ kg m^2
Dynamical ellipticity	$H = (C-A)/C = 3.273\,795(1) \times 10^{-3} = 1/305.457$
Ellipticity coefficient	$J_2 = \dfrac{C-A}{Ma^2} = 1.082\,626\,4(5) \times 10^{-3}$
Coefficient of moment of inertia	$\dfrac{J_2}{H} = \dfrac{C}{Ma^2} = 0.330\,698(2)$
Rotational angular velocity	$\omega = 7.292\,115 \times 10^{-5}$ rad s^{-1}
Sidereal day	86 164.10 s
Solar day	86 400 s
Sidereal year	$= 3.155\,815 \times 10^7$ s
Obliquity of ecliptic	$\theta = 23^\circ.4523$
Equatorial gravity (on geoid)	$g_e = 9.780\,319$ m s^{-2} (excluding atmosphere)
Ratio $\dfrac{\text{centrifugal force}}{\text{equatorial gravity}}$	$m = \dfrac{\omega^2 a}{g_e} = 3.467\,75 \times 10^{-3}$
Semi-major axis of orbit	$r_E = 1.495\,9789 \times 10^{11}$ m ($\equiv$ 1 AU)
Eccentricity of orbit	$e = 0.016\,73$

Table A.4 (cont.)

Date of perihelion (closest approach to Sun)	4 January
Mean orbital speed[b]	29 783.6 m s^{-1}
Orbital angular momentum	2.662×10^{40} kg m^2 s^{-1}
Ratio $\dfrac{\text{mass of Sun}}{\text{mass of Earth}}$	332 946.8
Solar constant	$S = 1372$ W m^{-2}
Mean Earth–Moon distance	$3.8440_5 \times 10^8$ m
Ratio $\dfrac{\text{mass of Earth}}{\text{mass of Moon}}$	$\mu = 81.300\,59$
Lunar orbital angular velocity	$\omega_L = 2.661\,698 \times 10^{-6}$ rad s^{-1}
Rate of precession of equinox	$\omega_p = 50''.291$ year^{-1} $= 7.7260 \times 10^{-12}$ rad s^{-1}
Period of precession	8.132×10^{11}s = 25 770 years
Total geothermal flux	$\dot{Q} = 4.42 \times 10^{13}$ W
Average geothermal flux	$\dot{Q}/A = 0.082$ W m^{-2}
Magnetic dipole moment (2005)	$m = 7.779 \times 10^{22}$ Am^2

[a] For an oblate ellipsoid, a > c, the surface area is $A = 2\pi a^2 + \dfrac{2\pi c^2}{\sqrt{1-\frac{c^2}{a^2}}} \ln\left[\dfrac{a}{c} + \sqrt{\dfrac{a^2}{c^2} - 1}\right]$.

[b] The circumference of an ellipse of semi-major axis a and eccentricity $e = (1 - b^2/a^2)^{1/2}$, where b is the semi-minor axis, is $4a\,\mathrm{E}(e) = 2\pi a\left(1 - \dfrac{1}{2^2}e^2 - \dfrac{1^2.3}{2^2.4^2}e^4 - \dfrac{1^2.3^2.5}{2^2.4^2.6^2}e^6 \ldots\right)$, where E($e$) is the complete elliptic integral of the second kind.

Table A.5 Summary of properties of the Sun and Moon

	Sun	Moon
Radius (m)	6.96×10^8	1.738×10^6
Mass (kg)	1.9884×10^{30}	7.3459×10^{22}
Mean density (kg m^{-3})	1408	3340.5
Central density (kg m^{-3})	1.6×10^5	~8000 (iron core)
Moment of inertia (kg m^2)	5.7×10^{46}	8.68×10^{34}
Mean distance from Earth (m)	$1.495\,9789 \times 10^{11}$	3.8440×10^8
Orbital angular velocity (rad s^{-1})	$1.990\,99 \times 10^{-7}$	$2.661\,70 \times 10^{-6}$
Eccentricity of orbit	0.01673	0.0549
Inclination of orbit to ecliptic (degrees)	0	5.14 ± 0.19 (variable)
Inclination of orbit to Earth's rotation (degrees)	23.438 63 (in 2005)	
Rotation rate (rad s^{-1})	2.87×10^{-6}	$2.661\,70 \times 10^{-6}$
Energy output (W)	3.846×10^{26}	~7×10^{11}

Table A.6 Physical properties of materials (values quoted are representative but some properties are very variable between samples)

	Granite	Basalt	Iron (20 °C)	Liquid iron (M P)	Sea water	Dry air (1 Atm, 15 °C)
Density, ρ (kg m^{-3})	2670	2900	7870	7010	1025(15 °C)[a]	1.226
Incompressibility, K_S (GPa)	55	67	170	130	2.05	1.42×10^{-4}
Rigidity, μ (GPa)	30	37	82	0	0	0
Viscosity, η (Pa s)	–	–	–	6×10^{-3}	1.7×10^{-3} (15 °C)	1.80×10^{-5}
Specific heat, C_p (J kg^{-1} K^{-1})	830	880	447	790	3990	1006
Vol. exp. coeff., α (10^{-6} K^{-1})	20	16	36	98	150(15 °C)	3480
Thermal cond., κ (W m^{-1}K^{-1})	3.0	2.5	75	36	0.59	0.0252
Thermal diff., η (10^{-6} m^2 s^{-1})	1.3_5	1.0	21	6.5	0.14	20.4
Latent heat of melting, L (10^5 J kg^{-1})	4.2	4.2	2.75	2.75	3.35	1.96

Melting point, T_M (K)	1440(S)
	1550(L)
		1350(S)
		1500(L)
			1812	1812	271.5	60

Elect. res., ρ_e (Ωm)	10^{10} (dry)	2×10^8 (dry)	0.098×10^{-6}	1.34×10^{-6}	0.23(15 °C)	5×10^{13} (at 100 Vm^{-1} in fine weather)
Dielectric const., k	8 (dry)	12 (dry)	–	–	80	$1 + 5.46 \times 10^{-4}$
Mag. susceptibility, $\chi_m = \left(\frac{\mu}{\mu_0} - 1\right)$	8×10^{-5}	2×10^{-3}	1000	1.8×10^{-5}	-9×10^{-9} (pure water)	3.74×10^{-7}

[a] Density of ice at 270 K = 917.5 kg m^{-3}

Table A.7 The Greek alphabet

Alpha	α	A	Nu	ν	N
Beta	β	B	Xi	ξ	Ξ
Gamma	γ	Γ	Omicron	o	O
Delta	δ	Δ	Pi	π	Π
Epsilon	ε	E	Rho	ρ	P
Zeta	ζ	Z	Sigma	σ	Σ
Eta	η	H	Tau	τ	T
Theta	θ	Θ	Upsilon	υ	Y
Iota	ι	I	Phi	ϕ	Φ
Kappa	κ	K	Chi	χ	X
Lambda	λ	Λ	Psi	ψ	Ψ

Appendix B

Orbital dynamics (Kepler's laws)

For many purposes it suffices to assume that planetary and satellite orbits are circular. Johannes Kepler (1571–1630) first recognized that they are elliptical from his analyses of observations by Tycho Brahe. The departure from a circle is expressed by the eccentricity

$$e = \left(1 - b^2/a^2\right)^{1/2}, \tag{B.1}$$

where a, b are the semi-major and semi-minor axes. Some asteroids have very elliptical orbits, but orbital eccentricities of the planets are mostly slight, being greatest for Pluto (0.250) and Mercury (0.2056). For the Earth's orbit, $e = 0.016\,73$.

Kepler summarized his conclusions in three empirical laws:

1. the orbit of each planet is an ellipse with the Sun at one focus;
2. the line between a planet and the Sun sweeps out equal areas in equal times;
3. the square of the orbital period of a planet is proportional to the cube of its mean distance from the Sun (which is equal to the semi-major axis of the orbit).

The second of these laws is a statement of the principle of conservation of angular momentum. The proof that the first and third laws are consequences of the inverse square law of gravitational attraction is the most widely known of Isaac Newton's discoveries.

A planetary orbit is confined to the plane defined by the instantaneous orbital velocity and the planet–Sun line. Since there is no velocity component or force on the planet perpendicular to this plane, the planet cannot leave it. Thus plane polar coordinates (r, θ) are most convenient to use in analysing the orbital problem, with the origin at the centre of mass of the Sun-plus-planet. The displacements of the planet, mass m, and Sun, mass M, from this origin remain in a fixed ratio, being inversely proportional to their masses, so that, whatever path a planet follows, the Sun follows its converse, appropriately diminished in size. If, at any instant, the planet is at a radial distance r (from the origin), the Sun is at a distance $(m/M)r$ in the opposite direction and the planet–Sun separation is $r(1 + m/M)$. The mutual attractive force is

$$F = \frac{GMm}{r^2(1 + m/M)^2}. \tag{B.2}$$

Thus the planet moves as though attracted to the coordinate origin by a mass fixed there, of magnitude

$$M' = M/(1 + m/M)^2. \tag{B.3}$$

In the case of the Earth and Sun, or, more correctly, (Earth plus Moon) and Sun, $m/M = 3 \times 10^{-6}$ and the orbital motion of the Sun is slight. The outer planets have bigger effects and, both directly and through their influence on the Sun, perturb the motion of the Earth from its simple two-body interaction with the Sun. In what follows, the motion of a planet, m, is analysed as though attracted to a fixed central mass M, acknowledging that this should, strictly, be M' by Eq. (B.3).

The orbital angular momentum of the planet, which is fixed, is

$$L = mr^2\omega = mr^2\frac{d\theta}{dt}, \tag{B.4}$$

where ω is angular velocity and θ is the angle of **r** with respect to a fixed direction in the orbital plane. But

$$r^2\frac{d\theta}{dt} = 2\frac{dS}{dt}, \tag{B.5}$$

where S is the area swept out by **r**. Thus

$$\frac{dS}{dt} = \frac{L}{2m}, \tag{B.6}$$

which is Kepler's second law.

The total planetary energy, the sum of kinetic and gravitational potential energies, is also conserved,

$$E = \frac{1}{2}mv^2 - \frac{GMm}{r} = \text{constant}, \tag{B.7}$$

where the planetary velocity, v, is the vector sum of radial and circumferential components,

$$\begin{aligned} v^2 &= \left(\frac{dr}{dt}\right)^2 + \left(r\frac{d\theta}{dt}\right)^2 = \left(\frac{dr}{d\theta}\frac{d\theta}{dt}\right)^2 + r^2\left(\frac{d\theta}{dt}\right)^2 \\ &= \left(\frac{d\theta}{dt}\right)^2\left[\left(\frac{dr}{d\theta}\right)^2 + r^2\right]. \end{aligned} \tag{B.8}$$

Substituting for (dθ/dt) in terms of L by Eq. (B.4) and for v^2 by Eq. (B.7) and rearranging, we obtain the differential equation for r(θ),

$$\frac{dr}{d\theta} = r\left(\frac{2mE}{L^2}r^2 + \frac{2GMm^2}{L^2}r - 1\right)^{1/2}. \tag{B.9}$$

As may be verified by differentiation and substitution, the solution is the equation for an ellipse with one focus at the origin,

$$r = \frac{p}{1 + e\cos(\theta + C)}, \tag{B.10}$$

where

$$p = \frac{L^2}{GMm^2} \tag{B.11}$$

and

$$(e^2 - 1) = \frac{2EL^2}{G^2M^2m^3}. \tag{B.12}$$

C is a constant of integration, being the phase angle of the orbit, which may be made zero by choosing the θ = 0 direction to coincide with the position of the planet at its closest approach to the Sun (perigee). That is at θ = 0,

$$r = r_p = r_{\min} = p/(1 + e). \tag{B.13}$$

Equation (B.10) demonstrates Kepler's first law.

The most remote point of an elliptical orbit (apogee) is, by Eq. (B.10),

$$r_a = p/(1 - e). \tag{B.14}$$

Combining this with Eq. (B.13) for the perigee distance, we obtain the semi-major axis,

$$a = 1/2(r_p + r_a) = p/(1 - e^2), \tag{B.15}$$

from which

$$r_p = a(1 - e), \tag{B.16}$$

$$r_a = a(1 + e), \tag{B.17}$$

and the centre of the ellipse is at distance ae from the focus (Sun). Using Eq. (B.1), we may also write the semi-minor axis in terms of p and e,

$$b = a\left(1 - e^2\right)^{1/2} = p/\left(1 - e^2\right)^{1/2} = (ap)^{1/2}. \tag{B.18}$$

The case of elliptical orbits is characterized by a negative value of total energy, given by Eq. (B.7), that is the energy is less than if the planet were at rest at infinite distance. It is held in orbit and by Eq. (B.12), $E < 0$ requires $e^2 < 1$, which corresponds to an ellipse. With $E = 0$ the planet would just escape, $e^2 = 1$ and the orbit would be parabolic. If $E > 0$, the planet would escape on a hyperbolic orbit ($e^2 > 1$). Another way of representing the energy is to combine Eqs. (B.11), (B.12) and (B.15) to obtain the simple result

$$E = -\frac{GMm}{2a}. \tag{B.19}$$

Thus, the total energy depends only on the semi-major axis of the orbit and is independent of its eccentricity. Combining this result with the fundamental equation for total energy (B.7), we obtain the *vis-viva* equation,

$$v^2 = GM\left(\frac{2}{r} - \frac{1}{a}\right), \tag{B.20}$$

which is a re-statement in a convenient form of the conservation of energy.

Kepler's third law is obtained by calculating the orbital area, S, because S is directly related to orbital period, T, by integrating Eq. (B.6),

$$S = \frac{L}{2m} T, \tag{B.21}$$

where

$$S = \int_0^{2\pi} \frac{1}{2} r^2 \mathrm{d}\theta = \frac{p^2}{2} \int_0^{2\pi} \frac{\mathrm{d}\theta}{(1 + e\cos\theta)^2}$$
$$= \frac{\pi p^2}{(1 - e^2)^{3/2}}, \tag{B.22}$$

which is a standard integral (see, for example, Dwight, 1961, item 858.535). Substituting for $(1 - e^2)$ and then p by Eqs. (B.15) and (B.11), we obtain

$$T^2 = \frac{4\pi^2}{GM} a^3, \tag{B.23}$$

which is the third law.

It is interesting to note that evidence from the Pioneer probes, presented by Anderson *et al.* (1998), indicates a possible breakdown in the inverse square law of gravity at the distances of the outer Solar System. It is not understood and is subject to further investigation. No allowance is made for it here. Also relativistic effects are neglected; these are just noticeable in the most precise satellite ranging.

Appendix C

Spherical harmonic functions

Spherical harmonic analysis may be regarded as the adaption of Fourier analysis to a spherical surface. It is therefore a convenient way of representing and analysing physical phenomena and properties that are distributed over the Earth's surface. However, spherical harmonics have a more fundamental significance than mere convenience. They are solutions of Laplace's equation, which is obeyed by potential fields (gravity, magnetism) outside the sources of the fields, and of the seismic wave equation in spherical geometry. Thus spherical harmonic representations are appropriate for the Earth's gravitational and magnetic fields and for free oscillations. Somewhat different procedures and normalizations are applied in the different sub-disciplines of geophysics. A statement of the mathematical properties of 'spherical harmonic functions is given in Chapter 3 of Sneddon (1980). Chapman and Bartels (1940, Vol.2) give details of the application to geomagnetism and the discussion by Kaula (1968) is useful, particularly in the application to gravity.

Laplace's equation is most familiar in Cartesian coordinates,

$$\nabla^2 V = \frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} + \frac{\partial^2 V}{\partial z^2} = 0. \tag{C.1}$$

Rewritten in spherical polar coordinates it is

$$\nabla^2 V = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 \frac{\partial V}{\partial r}\right) + \frac{1}{r^2 \sin\theta}\frac{\partial}{\partial \theta}\left(\sin\theta \frac{\partial V}{\partial \theta}\right) + \frac{1}{r^2 \sin^2\theta}\frac{\partial^2 V}{\partial \lambda^2} = 0, \tag{C.2}$$

where V is the potential describing a particular field. The origin ($r=0$) is normally the centre of the Earth. θ is the angle with respect to the chosen coordinate axis, commonly but not necessarily the Earth's rotational axis, in which case it is co-latitude ($90^\circ -$ latitude) and λ is longitude, measured from a convenient reference (the Greenwich meridian if not otherwise specified).

The wave equation can be written in similar form:

$$\frac{\partial^2 V}{\partial t^2} = c^2 \nabla^2 V, \tag{C.3}$$

where c is wave speed and V is the potential whose derivative in any direction gives the component of displacement in that direction. By imposing spherical geometry on this equation, we see that the solutions describing free oscillations of the Earth have the surface patterns of the spherical harmonic solutions of Laplace's equation (C.2).

Equation (C.2) is rendered tractable by assuming a separation of variables, that is V is the product of separate functions of r, θ and λ. This procedure is justified by the fact that it yields solutions of the form

$$V = \left[r^l, r^{-(l+1)}\right] \cdot [\cos m\lambda, \sin m\lambda] \cdot \mathrm{P}_{lm}(\cos\theta). \tag{C.4}$$

The square brackets give alternative solutions. l and m are integers with $m \le l$, and $\mathrm{P}_{lm}(\mu)$ satisfies the equation

$$\left(1-\mu^2\right)\frac{\mathrm{d}^2\mathrm{P}}{\mathrm{d}\mu^2} - 2\mu\frac{\mathrm{d}\mathrm{P}}{\mathrm{d}\mu} + \left[l(l+1) - \frac{m^2}{1-\mu^2}\right]\mathrm{P} = 0. \tag{C.5}$$

Table C.1 Legendre polynomials $P_l(\cos\theta)$ and associated polynomials, $p_{lm}(\cos\theta)$. Numerical factors convert P_{lm} to p_l^m

	$m=0$		$m=1$	
$l=0$	1	1	–	–
1	$\cos\theta$	$\sqrt{3}$	$\sin\theta$	$\sqrt{3}$
2	$\frac{1}{2}\left(3\cos^2\theta-1\right)$	$\sqrt{5}$	$3\cos\theta\sin\theta$	$\sqrt{5/3}$
3	$\frac{1}{2}\left(5\cos^3\theta-3\cos\theta\right)$	$\sqrt{7}$	$\frac{3}{2}\left(5\cos^2\theta-1\right)\sin\theta$	$\sqrt{7/6}$
4	$\frac{1}{8}\left(35\cos^4\theta-30\cos^2\theta+3\right)$	$\sqrt{9}$	$\frac{5}{2}\left(7\cos^3\theta-3\cos\theta\right)\sin\theta$	$\sqrt{9/10}$

	$m=2$		$m=3$		$m=4$	
$l=0$	–	–	–	–	–	–
1	–	–	–	–	–	–
2	$3\sin^2\theta$	$\sqrt{5/12}$	–	–	–	–
3	$15\cos\theta\sin^2\theta$	$\sqrt{7/60}$	$15\sin^3\theta$	$\sqrt{7/360}$	–	–
4	$\frac{15}{2}\left(7\cos^2\theta-1\right)\sin^2\theta$	$\sqrt{1/20}$	$105\cos\theta\sin^3\theta$	$\sqrt{1/280}$	$105\sin^4\theta$	$\sqrt{1/2240}$

This reduces to Legendre's equation for the case $m=0$, for which there is no variation in V with longitude. Equation (C.5) with $m\neq 0$ is Legendre's associated equation. Considering first the special case $m=0$, solutions of Eq. (C.5) have the form

$$\mathrm{P}_{l0}(\mu)=\frac{1}{2^l l!}\frac{\mathrm{d}^l}{\mathrm{d}\mu^l}\left[\left(\mu^2-1\right)^l\right], \tag{C.6}$$

where the multiplying factor $1/2^l l!$ does not affect the solution but normalizes the function so that $\mathrm{P}_{l0}(1)=+1$. Sneddon (1980) refers to this as Rodrigue's formula. The functions $\mathrm{P}_{l0}(\cos\theta)$, for the case $m=0$, are the Legendre polynomials, normally written with the second subscript omitted, $\mathrm{P}_l(\cos\theta)$. If convenient, latitude, ϕ, may be used instead of co-latitude, θ, by substituting $\mu=\cos\theta=\sin\phi$. Explicit expressions for the first few $\mathrm{P}_l(\cos\theta)$ are given in the $m=0$ column of Table C.1.

In putting $m=0$, we restrict the solutions to the description of potentials with rotational symmetry. These are zonal harmonics. As in Eq. (C.4), a potential that is expressed in zonal harmonics is written as a sum of terms, each of which is a power of r, with the Legendre polynomials appearing in the coefficients and representing the latitude variations. In geophysical problems it is convenient to make the coefficients dimensionally equal by normalizing r to the Earth's radius, a. Thus

$$V=\frac{1}{a}\sum_{l=0}^{\infty}\left[C_l\left(\frac{a}{r}\right)^{l+1}+C_l'\left(\frac{r}{a}\right)^l\right]\mathrm{P}_l(\cos\theta), \tag{C.7}$$

where the C_l are constant coefficients representing sources of potential inside the surface considered and the C_l' are due to external sources. Equation (C.7) is a sum of terms with the form of Eq. (C.4) with $m=0$.

The derivation of MacCullagh's formula for the gravitational potential due to a distributed mass with a slight departure from spherical symmetry can be extended to obtain a more complete solution, with the form of Eq. (C.7). If, instead of terminating the expansion of Eq. (6.4) at terms in $1/r^2$, we continue to higher powers in $1/r$, the coefficients are Legendre polynomials, i.e.

$$\left[1+\left(\frac{s}{r}\right)^2-2\frac{s}{r}\cos\psi\right]^{-1/2}=\sum_{l=0}^{\infty}\left(\frac{s}{r}\right)^l\mathrm{P}_l(\cos\psi). \tag{C.8}$$

Applying this expansion to the potential in Eq. (6.3), we obtain an infinite series with the form of Eq. (6.1) in which the coefficients J_l represent the multipole moments of the mass distribution.

Now consider the more general case of a potential that does not have rotational symmetry, so that $m \neq 0$ in Eqs. (C.4) and (C.5). By differentiation and substitution we can show that Eq. (C.5) has solutions

$$P_{lm}(\mu) = \left(1-\mu^2\right)^{m/2}\frac{\mathrm{d}^m}{\mathrm{d}\mu^m}[P_{l0}(\mu)]$$
$$= \frac{1}{2^l l!}\left(1-\mu^2\right)^{m/2}\frac{\mathrm{d}^{l+m}}{\mathrm{d}\mu^{l+m}}\left[\left(\mu^2-1\right)^l\right]. \qquad \text{(C.9)}$$

These are the associated Legendre polynomials, or Ferrer's modified version (Sneddon, 1980), introduced to avoid factors $(-1)^{m/2}$ in the original. It is a straightforward matter to calculate the first few functions directly from Eq. (C.9), but a more convenient polynomial form is

$$P_{lm}(\cos\theta) = \frac{\sin^m\theta}{2^l}\sum_{t=0}^{\mathrm{Int}[(l-m)/2]} \frac{(-1)^t(2l-2t)!}{t!(l-t)!\,(l-m-2t)!}\cos^{l-m-2t}\theta. \qquad \text{(C.10)}$$

The upper limit of this summation is the integral part of $[(l-m)/2]$, ignoring the extra 1/2 for odd values of $(l-m)$. Explicit forms for the lowest degrees (l) and orders (m) are listed in Table C.1.

The general expression for potential as a sum of spherical harmonics is

$$V = -\frac{1}{a}\sum_{l=0}^{\infty}\sum_{m=0}^{l}\left\{\begin{matrix}\left[C_{lm}\left(\frac{a}{r}\right)^{l+1}+C'_{lm}\left(\frac{r}{a}\right)^{l}\right]\cos m\lambda \\ +\left[S_{lm}\left(\frac{a}{r}\right)^{l+1}+S'_{lm}\left(\frac{r}{a}\right)^{l}\right]\sin m\lambda\end{matrix}\right\} \times P_{lm}(\cos\theta), \qquad \text{(C.11)}$$

which is simply a sum of terms with the form of Eq. (C.4). As with Eq. (C.7) the unprimed coefficients refer to internal sources, which have vanishing influence at $r \to \infty$, and the primed coefficients are attributable to external sources.

The general surface patterns of spherical harmonics can be seen by considering Eqs. (C.4) or (C.11) and (C.10). Around any complete (360°) line of latitude (at fixed θ) there is a sinusoidal variation with λ, crossing $2m$ meridians where the function vanishes. The latitude variation is less obvious, but examination of Eq. (C.10) shows that down any selected meridian, that is over 180° from pole to pole, there are $(l-m)$ values of latitude where the function vanishes. Thus l gives the total number of nodal lines on one hemisphere and is a measure of the fineness of the structure represented; m determines the distribution of the total between lines of latitude and longitude (Fig. C.1). For $m=0$ they are all latitudinal and for $m=l$ they are all longitudinal. The upper limit of m is l, as can be seen in Eq. (C.9) because when $(\mu^2-1)^l$ is differentiated $(l+m)$ times the derivative vanishes for $m>l$.

For the purposes of this text we are interested mainly in the terms with unprimed coefficients in Eq. (C.11). These terms all decrease with increasing distance from the origin (and from the internal source of the potential) and the rate of decrease increases with l. The low harmonic degrees become increasingly dominant at greater distances. Conversely, extrapolating downwards from surface observations, the higher harmonic degrees become increasingly prominent. The harmonics are expressing the obvious principle that fine details at depth are difficult to discern at the surface because of the spatial attenuation.

A common feature of the terms in a Fourier series and the Legendre and associated polynomials is that they are orthogonal. This means

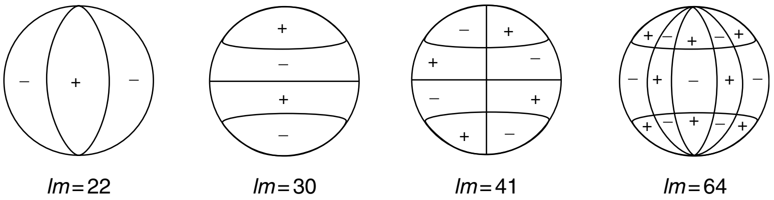


FIGURE C.1. Examples of spherical harmonics. $m=0$ gives zonal harmonics, $m=l$ gives sectoral harmonics and general cases, $0<m<l$, are known as tesseral harmonics.

that the integral over a sphere of any product vanishes,

$$\int_0^{2\pi}\int_{-1}^{1} \mathrm{P}_{lm}(\mu)\cdot \mathrm{P}_{l'm'}(\mu)\cdot[\cos m\lambda, \sin m\lambda]. \\ [\cos m'\lambda, \sin m'\lambda]\mathrm{d}\mu\ \mathrm{d}\lambda = 0, \qquad (C.12)$$

unless both $l' = l$ and $m' = m$. This means that in a harmonic analysis of a complete data set over a spherical surface the coefficients of the harmonic series are independent and errors are not introduced by truncating the series. However, the calculated coefficients are affected by truncation when discrete or irregularly spaced data are used, as is normally the case.

In some treatments of Legendre polynomials the subscript n is used in place of l. Here n is reserved for a further development that appears in the study of free oscillations – a harmonic radial variation. Free oscillations are classified in Section 5.3 according to the values of three integers; thus ${}_n\mathrm{S}_l^m$ and ${}_n\mathrm{T}_l^m$ denote spheroidal and torsional oscillations, respectively, where l, m represent variations on a spherical surface, as for $\mathrm{P}_{lm}(\cos\theta)\cos m\lambda$, and n is the number of internal spherical surfaces that are nodes of the motion.

The numerical factors in the associated polynomials defined by Eqs. (C.9) and (C.10) increase rapidly with m; to make the coefficients in a harmonic analysis relate more nearly to the physical significance of the terms they represent, various normalizing factors are used. The one that has been employed in most recent analyses of the geoid and must be favoured for general adoption is the 'fully normalized' function

$$\mathrm{p}_l^m(\cos\theta) = \left[(2-\delta_{m,0})(2l+1)\frac{(l-m)!}{(l+m)!}\right]^{1/2}\mathrm{P}_{lm}(\cos\theta), \qquad (C.13)$$

which is so defined that

$$\frac{1}{4\pi}\int_0^{2\pi}\int_{-1}^{1}\{\mathrm{p}_l^m(\cos\theta)[\sin m\lambda, \cos m\lambda]\}^2\mathrm{d}(\cos\theta)\mathrm{d}\lambda = 1, \qquad (C.14)$$

that is, the mean square value over a spherical surface is unity. Note that the factor $(2-\delta_{m,o})$ in Eq. (C.12) is unity if $m = 0$, but $(2-\delta_{m,o}) = 2$ if $m \neq 0$ because the factor $[\sin m\lambda, \cos m\lambda]^2$ in Eq. (C.14) introduces a factor 1/2. (There is no alternative $\sin m\lambda$ term if $m = 0$.) The coefficients of a spherical harmonic expansion referred to the normalized coefficients, p_l^m, are distinguished by a bar: $\overline{C}_l^m, \overline{S}_l^m$. Thus

$$V = \frac{1}{a}\sum_{l=0}^{\infty}\sum_{m=0}^{l}\left\{\begin{array}{l}\left[\overline{C}_l^m\left(\frac{a}{r}\right)^{l+1} + \overline{C'}_l^m\left(\frac{a}{r}\right)^{l}\right]\cos m\lambda \\ +\left[\overline{S}_l^m\left(\frac{a}{r}\right)^{l+1} + \overline{S'}_l^m\left(\frac{r}{a}\right)^{l}\right]\sin m\lambda\end{array}\right\} \\ \times\,\mathrm{p}_l^m(\cos\theta). \qquad (C.15)$$

The spherical harmonics used in geomagnetism apply no normalization to the zonal harmonics ($m = 0$), but bring the sectoral and tesseral harmonics into line with the zonal harmonics of the same degree by applying the normalizing factor $\left[(2-\delta_{m,0})(l-m)!/(l+m)!\right]^{1/2}$.

It is of interest to consider the spherical harmonic expansions of some simple surface patterns that are subject to analytical representation. Thus, if we consider the equation for the surface of an oblate ellipsoid of equatorial radius a and eccentricity e,

$$r = a\left(1 + \frac{e^2}{1-e^2}\sin^2\phi\right)^{-1/2}, \qquad (C.16)$$

and expand in powers of e to e^6 or flatten $f = (1 - c/a)$ to f^3 and zonal harmonics to P_6, we have

$$\frac{r}{a} = \left(1 - \frac{e^2}{6} - \frac{11}{20}e^4 - \frac{103}{1680}e^6\right) \\ + \left(-\frac{e^2}{3} - \frac{5}{42}e^4 - \frac{3}{56}e^6\right)\mathrm{P}_2 \\ + \left(\frac{3}{35}e^4 + \frac{57}{770}e^6\right)\mathrm{P}_4 - \frac{5}{231}e^6\mathrm{P}_6, \qquad (C.17)$$

$$\frac{r}{a} = \left(1 - \frac{f}{3} - \frac{f^2}{5} - \frac{13}{105}f^3\right) \\ + \left(-\frac{2}{3}f - \frac{1}{7}f^2 + \frac{1}{21}f^3\right)\mathrm{P}_2 \\ + \left(\frac{12}{35}f^2 + \frac{96}{385}f^3\right)\mathrm{P}_4 - \frac{40}{231}f^3\mathrm{P}_6. \qquad (C.18)$$

Note that the P_2 term alone does not represent an ellipsoidal surface, although for the Earth the ellipticity is sufficiently slight that expansion to e^4, f^2, P_4 suffices.

If we consider a single surface spike of negligible lateral dimensions, we obtain equal coefficients for all unnormalized zonal harmonics or amplitudes proportional to $(2l+1)^{-1/2}$ in fully normalized harmonics. For a pair of opposite points the even terms follow the same pattern, but the odd terms vanish. Another geometrically simple case is a great circle line source, which also gives vanishing odd terms, but even coefficients (fully normalized) varying as

$$\overline{C_l} = (-1)^{l/2} 1/2^l \left[(l/2)!\right]^2 (2l+1)^{1/2}. \tag{C.19}$$

The values of this function oscillate in sign with amplitudes very close to $1.12/(2l+1)$, except $C_0 = 1$.

Appendix D

Relationships between elastic moduli of an isotropic solid

Table D.1 Elastic moduli (see Section 10.2)

	K	μ	ν	**E**	λ	χ
K,μ	K	μ	$\dfrac{3K-2\mu}{6K+2\mu}$	$\dfrac{9K\mu}{3K+\mu}$	$K-\dfrac{2}{3}\mu$	$K+\dfrac{4}{3}\mu$
K,ν	K	$\dfrac{3K(1-2\nu)}{2(1+\nu)}$	ν	$3K(1-2\nu)$	$\dfrac{3K\nu}{1+\nu}$	$\dfrac{3K(1-\nu)}{(1+\nu)}$
K,E	K	$\dfrac{3KE}{9K-E}$	$\dfrac{1}{2}-\dfrac{E}{6K}$	E	$\dfrac{3K(3K-E)}{9K-E}$	$\dfrac{3K(3K+E)}{9K-E}$
K,λ	K	$\dfrac{3}{2}(K-\lambda)$	$\dfrac{\lambda}{3K-\lambda}$	$\dfrac{9K(K-\lambda)}{3K-\lambda}$	λ	$3K-2\lambda$
K,χ	K	$\dfrac{3}{4}(\chi-K)$	$\dfrac{3K-\chi}{3K+\chi}$	$\dfrac{9K(\chi-K)}{3K+\chi}$	$\dfrac{1}{2}(3K-\chi)$	χ
μ,ν	$\dfrac{2\mu(1+\nu)}{3(1-2\nu)}$	μ	ν	$2\mu(1+\nu)$	$\dfrac{2\mu\nu}{1-2\nu}$	$\dfrac{2\mu(1-\nu)}{(1-2\nu)}$
μ,E	$\dfrac{\mu E}{3(3\mu-E)}$	μ	$\dfrac{E}{2\mu}-1$	E	$\dfrac{\mu(E-2\mu)}{3\mu-E}$	$\dfrac{\mu(4\mu-E)}{3\mu-E}$
μ,λ	$\lambda+\dfrac{2}{3}\mu$	μ	$\dfrac{\lambda}{2(\lambda+\mu)}$	$\dfrac{\mu(3\lambda+2\mu)}{\lambda+\mu}$	λ	$\lambda+2\mu$
μ,χ	$\chi-\dfrac{4}{3}\mu$	μ	$\dfrac{\chi-2\mu}{2(\chi-\mu)}$	$\dfrac{\mu(3\chi-4\mu)}{\chi-\mu}$	$\chi-2\mu$	χ
ν,E	$\dfrac{E}{3(1-2\nu)}$	$\dfrac{E}{2(1+\nu)}$	ν	E	$\dfrac{E\nu}{(1+\nu)(1-2\nu)}$	$\dfrac{E(1-\nu)}{(1+\nu)(1-2\nu)}$
ν,λ	$\dfrac{\lambda(1+\nu)}{3\nu}$	$\dfrac{\lambda(1-2\nu)}{2\nu}$	ν	$\dfrac{\lambda(1+\nu)(1-2\nu)}{\nu}$	λ	$\dfrac{\lambda(1-\nu)}{\nu}$
ν,χ	$\dfrac{\chi(1+\nu)}{3(1-\nu)}$	$\dfrac{\chi(1-2\nu)}{2(1-\nu)}$	ν	$\dfrac{\chi(1+\nu)(1-2\nu)}{1-\nu}$	$\dfrac{\chi\nu}{1-\nu}$	χ

Table D.1 (cont.)

	K	μ	ν	E	λ	χ
E,λ	$\dfrac{E+3\lambda+p}{6}$	$\dfrac{E-3\lambda+p}{4}$	$\dfrac{p-E-\lambda}{4\lambda}$	E	λ	$\dfrac{E-\lambda+p}{2}$
E,χ	$\dfrac{3\chi-E+q}{6}$	$\dfrac{E+3\chi-q}{8}$	$\dfrac{E-\chi+q}{4\chi}$	E	$\dfrac{\chi-E+q}{4}$	χ
λ,χ	$\dfrac{1}{3}(2\lambda+\chi)$	$\dfrac{1}{2}(\chi-\lambda)$	$\dfrac{\lambda}{\lambda+\chi}$	$\dfrac{(2\lambda+\chi)(\chi-\lambda)}{\lambda+\chi}$	λ	χ

$p=\sqrt{E^2+2E\lambda+9\lambda^2}$; $q=\sqrt{E^2-10E\chi+9\chi^2}$

Note: for mathematical convenience it is sometimes assumed that there is only one independent modulus, with $\lambda=\mu$ and therefore $K=5\mu/3$, $\chi=3\mu=9K/5$, $\nu=1/4$. This is referred to as a Poisson solid, but it is not a good approximation for rocks and becomes increasingly unsatisfactory with increasing pressure.

Appendix E

Thermodynamic parameters and derivative relationships

Tables E.2 and E.3 present a compact summary of thermodynamic derivatives in a form convenient for geophysical applications. Individual entries have no meaning; they must be taken in pairs so that, for example, to find $(\partial T/\partial P)_S$ look down the constant S column and take the ratio of entries for ∂T and ∂P, that is $\gamma T/K_S$. An arbitrary mass m of material is assumed, so that m appears in many of the entries. Table E.2 is complete for the eight primary parameters. Any one of them may be differentiated with respect to any other one with any third one held constant. The results are represented in terms of the same parameters plus a set of first derivative properties, α, K_T, K_S, C_V, C_P and γ. Table E.3 extends the constant T, P, V and S columns to derivatives of these first derivative properties, using a set of second derivative parameters, K_T', K_S', δ_T, δ_S, C_T', C_S' and q, defined in Table E.1. There are numerous alternative forms for many of the Table E.3 entries, the usefulness of which depends on particular applications. Substitutions may be made using relationships in Table E.4. A few derivatives of products and third derivatives have been found useful and are also listed.

The compact collection of thermodynamic derivatives in Tables E.2 and E.3 follows an idea, started by Bridgman (1914), that is much more useful than generally realized. Bridgman's original compilation was difficult to use because it related derivatives to one another and not to familiar parameters, as in the tables presented here. Also, some confusion arose from errors that were copied in later compilations and not corrected for another 80 years (Dearden, 1995).

Table E.1 Thermodynamic notation and definitions (note that parameters V, S, U, H, F and G refer to arbitrary mass, m, but C_V and C_P refer to unit mass)

Specific heat, constant *P*	$C_P = (T/m)(\partial S/\partial T)_P$
constant *V*	$C_V = (T/m)(\partial S/\partial T)_V$
	$C_S' = (\partial \ln C_V/\partial \ln V)_S$; $C_T' = (\partial \ln C_V/\partial \ln V)_T$
Helmholtz free energy	$F = U - TS$
Gibbs free energy	$G = U - TS + PV$
Enthalpy	$H = U + PV$
Bulk modulus, adiabatic	$K_S = -V(\partial P/\partial V)_S$
	$K_S' = (\partial K_S/\partial P)_S$; $K_S'' = (\partial K_S'/\partial P)_S$
isothermal	$K_T = -V(\partial P/\partial V)_T$
	$K_T' = (\partial K_T/\partial P)_T$; $K_T'' = (\partial K_T'/\partial P)_T$
Pressure	P
	$q = (\partial \ln \gamma/\partial \ln V)_T = (\partial \ln(\gamma C_V)/\partial \ln V)_S$
	$q_S = (\partial \ln \gamma/\partial \ln V)_S = q - C_S'$
Heat	Q
Entropy	$S = \int \mathrm{d}Q/T$
Temperature	T
Internal energy	U
Volume	V
Volume expansion coefficient	$\alpha = (1/V)(\partial V/\partial T)_P$
Grüneisen parameter	$\gamma = \alpha K_T/\rho C_V = \alpha K_S/\rho C_P$
Anderson–Grüneisen parameter, adiabatic	$\delta_S = -(1/\alpha)(\partial \ln K_S/\partial T)_P = (\partial \ln(\alpha T/C_P)/\partial \ln V)_S$
isothermal	$\delta_T = -(1/\alpha)(\partial \ln K_T/\partial T)_P = (\partial \ln \alpha/\partial \ln V)_T$
Density	$\rho = m/V$
	$\lambda = (\partial \ln q/\partial \ln V)_T$

Table E.2 First order derivatives of thermodynamic parameters

Differential	Constant							
element	T	P	V	S	U	H	F	G
∂T	–	1	1	γT	$P - \alpha K_T T$	$1 - \alpha T$	P	1
∂P	$-K_T/V$	–	$\alpha K_T = \gamma \rho C_V$	K_S	$-\rho C_V(K_S - \gamma P)$	$-\rho C_P$	$K_T(S/V + \alpha P)$	S/V
∂V	1	αV	–	$-V$	mC_V	$\alpha V(1 + 1/\gamma)$	$-S$	$\alpha V - S/K_T$
∂S	$\alpha K_T = \gamma \rho C_V$	mC_P/T	mC_V/T	–	$mC_V P/T$	mC_P/T	$mC_V(P/T - \gamma S/V)$	$mC_P/T - \alpha S$
∂U	$\alpha K_T T - P$	$mC_P - \alpha VP$	mC_V	PV	–	$mC_P - PV\alpha$ $(1 + 1/\gamma)$	$mC_V P - S\alpha K_T T$ $+ SP$	$mC_P - \alpha TS - P\alpha V$ $+ SP/K_T$
∂H	$-K_T(1 - \alpha T)$	mC_P	$mC_V(1 + \gamma)$	$K_S V$	$mC_V[P(1 + \gamma)$ $- K_S]$	–	$SK_T(1 - \alpha T)$ $+ mC_V P(1 + \gamma)$	mC_P $+ S(1 - \alpha T)$
∂F	$-P$	$-S - \alpha VP$	$-S$	$PV - \gamma TS$	$\rho C_V(\gamma TS - PV)$ $- PS$	$-S(1 - \alpha T)$ $- PV\alpha(1 + 1/\gamma)$	–	$-S(1 - P/K_T)$ $- P\alpha V$
∂G	$-K_T$	$-S$	$-S + \alpha K_T V$	$K_S V - \gamma TS$	$mC_V(\gamma TS/V$ $+ \gamma P - K_S) - PS$	$-S(1 - \alpha T)$ $- mC_P$	$S(K_T - P)$ $+ PV\alpha K_T$	–

Table E.3 Thermodynamic derivatives extended to second order at constant *T*, *P*, *V* and *S*

Differential element	Constant			
	T	P	V	S
∂T	–	1	1	γT
∂P	$-K_T/V$	–	$\alpha K_T = \gamma\rho C_V$	K_S
∂V	1	αV	–	$-V$
∂S	$\alpha K_T = \gamma\rho C_V$	mC_P/T	mC_V/T	–
∂U	$\alpha K_T T - P$	$mC_P - \alpha VP$	mC_V	PV
∂H	$-K_T(1-\alpha T)$	mC_P	$mC_V(1+\gamma)$	$K_S V$
∂F	$-P$	$-S-\alpha VP$	$-S$	$PV-\gamma TS$
∂G	$-K_T$	$-S$	$-S+\alpha K_T V$	$K_S V - \gamma TS$
$\partial\alpha$	$\alpha\delta_T/V = -(\partial K_T/\partial T)_P/K_T V$	$\alpha^2(2\delta_T - K'_T + C'_T/\gamma\alpha T)$	$\alpha^2(\delta_T - K'_T + C'_T/\gamma\alpha T)$	$-\alpha[K'_S - 1 + q + \gamma\alpha T(\delta_S + q)]$
∂K_T	$-K_T K'_T/V$	$-\alpha K_T\delta_T = K_T{}^2(\partial\alpha/\partial P)_T$	$\alpha K_T(K'_T - \delta_T)$	$K_T[K'_T + \gamma\alpha T(K'_T - \delta_T)]$
∂K_S	$(-K_T/V)(K'_S + \gamma\alpha T\delta_S)$	$-\alpha K_S\delta_S$	$\alpha K_T(K'_S - \delta_S)$	$K_S K'_S$
∂C_V	$(C_V/V)C'_T = (C_V/V)(1-q+\delta_T - K'_T)$	$(C_P C'_T - C_V C'_S)/\gamma T$	$(C_V/\gamma T)(C'_T - C'_S)$	$-TC_V(\partial\gamma\partial T)_V$
∂C_P	$(C_P/V)[C'_T + \gamma\alpha T(q+\delta_T)/(1+\gamma\alpha T)]$	$(C_P/\gamma T)[C'_T - C'_S + \gamma\alpha T$ $\times(\delta_T - \delta_S + C'_T)]$	$(C_P/\gamma T)\{C'_T(1+\gamma\alpha T) + [\gamma^2\alpha T + (\gamma\alpha T)^2$ $\times(q-1) - C'_S]/(1+\gamma\alpha T)\}$	$-C_P[T(\partial\gamma/\partial T)_V + \gamma\alpha T(\delta_S + q)]$
$\partial\gamma$	$\gamma q/V$	$\gamma\alpha q + C'_S/T$	C'_S/T	$-\gamma(q - C'_S)$

Table E.4 Relationships between derivatives

$$K_S/K_T = C_P/C_V = 1 + \gamma\alpha T \tag{E.1}$$

$$K'_T = K'_S(1 + \gamma\alpha T) + \gamma\alpha T[3q - 2 - \gamma + \gamma(\partial \ln C_V/\partial \ln T)_V] \tag{E.2}$$

$$K'_S = K'_T(1 + \gamma\alpha T) - \gamma\alpha T(\delta_S + \delta_T + q) \tag{E.3}$$

$$\delta_S = -(1/\alpha)(\partial \ln K_S/\partial T)_P = K'_S - 1 + q - \gamma - C'_S = (\partial \ln(\alpha T/C_P)/\partial \ln V)_S \tag{E.4}$$

$$\begin{aligned} \delta_T &= -(1/\alpha)(\partial \ln K_T/\partial T)_P = K'_T - 1 + q + C'_T = (\partial \ln \alpha/\partial \ln V)_T \\ &= (\delta_S + C'_T)(1 + \gamma\alpha T) + \gamma + C'_S + \gamma\alpha T(2q - 1) \end{aligned} \tag{E.5}$$

$$C'_S = C'_T - \gamma(\partial \ln C_V/\partial \ln T)_V = (\partial\gamma/\partial \ln T)_V \tag{E.6}$$

$$C'_T = \gamma(\partial \ln(\gamma C_V)/\partial \ln T)_V \tag{E.7}$$

$$q_S = q - C'_S \tag{E.8}$$

$$(\partial \ln(\alpha K_T)/\partial \ln V)_T = \delta_T - K'_T = -(1/\alpha)(\partial \ln K_T/\partial T)_V \tag{E.9}$$

$$(\partial \ln(\alpha K_T)/\partial \ln T)_V = (\partial \ln(\gamma C_V)/\partial \ln T)_V = C'_T/\gamma \tag{E.10}$$

$$(\partial(\alpha K_T)/\partial T)_P = K_T(\partial\alpha/\partial T)_V \tag{E.11}$$

$$(\partial \ln(\alpha K_T)/\partial \ln V)_S = q - 1 \tag{E.12}$$

$$(\partial \ln(\alpha K_S)/\partial \ln V)_S = q - 1 + \gamma\alpha T(\delta_S + q) \tag{E.13}$$

$$(\partial \ln(\gamma\alpha T)/\partial \ln V)_T = \delta_T + q \tag{E.14}$$

$$(\partial \ln(\gamma\alpha T)/\partial \ln V)_S = (1 + \gamma\alpha T)(\delta_S + q) \tag{E.15}$$

$$(\partial K'_T/\partial T)_P = \alpha\delta_T[\delta_T - K'_T + (\partial \ln \delta_T/\partial \ln V)_T] \tag{E.16}$$

$$(\partial K'_S/\partial T)_P = \alpha\delta_S[\delta_S - K'_S + (\partial \ln \delta_S/\partial \ln V)_S] \tag{E.17}$$

$$(\partial\delta_T/\partial \ln V)_T = -K_T K''_T + \lambda q + (\partial C'_T/\partial \ln V)_T \tag{E.18}$$

$$(\partial\delta_S/\partial \ln V)_S = -K_S K''_S - \gamma q_S + (\partial q_S/\partial \ln V)_S \tag{E.19}$$

$$(\partial C_P/\partial P)_T = -(\partial(\alpha/\rho)/\partial \ln T)_P \tag{E.20}$$

Appendix F

An Earth model: mechanical properties

Table F.1 Selected details of the Preliminary Reference Earth Model (PREM) by Dziewonski and Anderson (1981)

Region	Radius (km)	V_P (m s^{-1})	V_S (m s^{-1})	ρ (kg m^{-3})	K_S (GPa)	μ (GPa)	ν	P (GPa)	g (m s^{-2})
Inner core	0	11 266.20	3667.80	13 088.48	1425.3	176.1	0.4407	363.85	0
	200	11 255.93	3663.42	13 079.77	1423.1	175.5	0.4408	362.90	0.7311
	400	11 237.12	3650.27	13 053.64	1416.4	173.9	0.4410	360.03	1.4604
	600	11 205.76	3628.35	13 010.09	1405.3	171.3	0.4414	355.28	2.1862
	800	11 161.86	3597.67	12 949.12	1389.8	167.6	0.4420	348.67	2.9068
	1000	11 105.42	3558.23	12 870.73	1370.1	163.0	0.4428	340.24	3.6203
	1200	11 036.43	3510.02	12 774.93	1346.2	157.4	0.4437	330.05	4.3251
	1221.5	11 028.27	3504.32	12 763.60	1343.4	156.7	0.4438	328.85	4.4002
Outer core	1221.5	10 355.68	0	12 166.34	1304.7	0	0.5	328.85	4.4002
	1400	10 249.59	0	12 069.24	1267.9	0	0.5	318.75	4.9413
	1600	10 122.91	0	11 946.82	1224.2	0	0.5	306.15	5.5548
	1800	9985.54	0	11 809.00	1177.5	0	0.5	292.22	6.1669
	2000	9834.96	0	11 654.78	1127.3	0	0.5	277.04	6.7715
	2200	9668.65	0	11 483.11	1073.5	0	0.5	260.68	7.3645
	2400	9484.09	0	11 292.98	1015.8	0	0.5	243.25	7.9425
	2600	9278.76	0	11 083.35	954.2	0	0.5	224.85	8.5023
	2800	9050.15	0	10 853.21	888.9	0	0.5	205.60	9.0414
	3000	8795.73	0	10 601.52	820.2	0	0.5	185.64	9.5570
	3200	8512.98	0	10 327.26	748.4	0	0.5	165.12	10.0464
	3400	8199.39	0	10 029.40	674.3	0	0.5	144.19	10.5065
	3480	8064.82	0	9903.49	644.1	0	0.5	135.75	10.6823
D″	3480	13 716.60	7264.66	5566.45	655.6	293.8	0.3051	135.75	10.6823
	3600	13 687.53	7265.75	5506.42	644.0	290.7	0.3038	128.71	10.5204
	3630	13 680.41	7265.97	5491.45	641.2	289.9	0.3035	126.97	10.4844
Lower mantle	3630	13 680.41	7265.97	5491.45	641.2	289.9	0.3035	126.97	10.4844
	3800	13 447.42	7188.92	5406.81	609.5	279.4	0.3012	117.35	10.3095

Table F.1 (cont.)

Region	Radius (km)	$V_P\,(\mathrm{m\,s^{-1}})$	$V_S\,(\mathrm{m\,s^{-1}})$	$\rho\,(\mathrm{kg\,m^{-3}})$	K_S (GPa)	μ (GPa)	ν	P (GPa)	$g\,(\mathrm{m\,s^{-2}})$
	4000	13 245.32	7099.74	5307.24	574.4	267.5	0.2984	106.39	10.1580
	4200	13 015.79	7010.53	5207.13	540.9	255.9	0.2957	95.76	10.0535
	4400	12 783.89	6919.57	5105.90	508.5	244.5	0.2928	85.43	9.9859
	4600	12 544.66	6825.12	5002.99	476.6	233.1	0.2898	75.36	9.9474
	4800	12 293.16	6725.48	4897.83	444.8	221.5	0.2864	65.52	9.9314
	5000	12 024.45	6618.91	4789.83	412.8	209.8	0.2826	55.9	9.9326
	5200	11 733.57	6563.70	4678.44	380.3	197.9	0.2783	46.49	9.9467
	5400	11 415.60	6378.13	4563.07	347.1	185.6	0.2731	37.29	9.9698
------------	5600	11 065.57	6240.46	4443.17	313.3	173.0	0.2668	28.29	9.9985
	5600	11 065.57	6240.46	4443.17	313.3	173.0	0.2668	28.29	9.9985
	5701	10 751.31	5945.08	4380.71	299.9	154.8	0.2798	23.83	10.0143
Transition zone	5701	10 266.22	5570.20	3992.14	255.6	123.9	0.2914	23.83	10.0143
------------	5771	10 157.82	5516.01	3975.84	248.9	121.0	0.2909	21.04	10.0038
	5771	10 157.82	5516.01	3975.84	248.9	121.0	0.2909	21.04	10.0038
	5871	9645.88	5224.28	3849.80	218.1	105.1	0.2924	17.13	9.9883
------------	5971	9133.97	4932.59	3723.78	189.9	90.6	0.2942	13.35	9.9686
	5971	8905.22	4769.89	3543.25	173.5	80.6	0.2988	13.35	9.9686
	6061	8732.09	4706.90	3489.51	163.0	77.3	0.2952	10.20	9.9361
	6151	8558.96	4643.91	3435.78	152.9	74.1	0.2914	7.11	9.9048
Low velocity zone	6151	7989.70	4418.85	3359.50	127.0	65.6	0.2797	7.11	9.9048
	6221	8033.70	4443.61	3367.10	128.7	66.5	0.2796	4.78	9.8783
	6291	8076.88	4469.53	3374.71	130.3	67.4	0.2793	2.45	9.8553
LID	6291	8076.88	4469.53	3374.71	130.3	67.4	0.2793	2.45	9.8553
	6346.6	8110.61	4490.94	3380.76	131.5	68.2	0.2789	0.604	9.8394
Crust	6346.6	6800.00	3900.00	2900.00	75.3	44.1	0.2549	0.604	9.8394
------------	6356	6800.00	3900.00	2900.00	75.3	44.1	0.2549	0.337	9.8332
	6356	5800.00	3200.00	2600.00	52.0	26.6	0.2812	0.337	9.8332
------------	6368	5800.00	3200.00	2600.00	52.0	26.6	0.2812	0.030	9.8222
Ocean	6368	1450.00	0	1020.00	2.1	0	0.5	0.030	9.8222
	6371	1450.00	0	1020.00	2.1	0	0.5	0	9.8156

Table F.2 Elastic properties of the core from equation of state fitting to PREM

r (km)	P (GPa)	K_S (GPa)	K'	KK''	μ (GPa)	μ'	$K\mu''$	ρ (kg m^{-3})
0	363.85	1444.03	3.3203	– 0.7118	174.04	0.1996	– 0.3061	13 082.19
200	362.90	1440.83	3.3207	– 0.7129	173.86	0.1997	– 0.3066	13 073.59
400	360.03	1431.17	3.3220	– 0.7165	173.23	0.2003	– 0.3081	13 047.66
600	355.28	1415.19	3.3243	– 0.7225	172.19	0.2013	– 0.3107	13 004.66
800	348.67	1392.98	3.3275	– 0.7311	170.76	0.2027	– 0.3144	12 944.53
1000	340.24	1364.66	3.3318	– 0.7427	168.94	0.2045	– 0.3193	12 867.21
1200	330.05	1330.43	3.3373	– 0.7574	166.73	0.2069	– 0.3257	12 772.67
1221.5	328.85	1326.36	3.3379	– 0.7691	166.46	0.2071	– 0.3264	12 761.17
1221.5	328.85	1301.35	3.3171	– 0.7034				12 163.35
1400	318.75	1267.81	3.3227	– 0.7182				12 068.11
1600	306.15	1225.90	3.3300	– 0.7378				11 946.76
1800	292.22	1179.46	3.3387	– 0.7612				11 809.17
2000	277.04	1128.70	3.3489	– 0.7889				11 654.83
2200	260.68	1073.81	3.3609	– 0.8182				11 482.94
2400	243.25	1015.11	3.3749	– 0.8609				11 292.85
2600	224.85	952.87	3.3914	– 0.9077				11 083.58
2800	205.60	887.40	3.4110	– 0.9641				10 855.02
3000	185.64	818.09	3.4344	– 1.0329				10 602.94
3200	165.12	748.33	3.4625	– 1.1180				10 328.76
3400	144.19	675.51	3.4970	– 1.2253				10 029.29
3480	135.75	645.93	3.5129	– 1.2762				9901.97

Table F.3 Elastic properties of the lower mantle from equation of state fitting to PREM

r (km)	P (GPa)	K_S (GPa)	K'	KK''	μ (GPa)	μ'	$K\mu''$	ρ (kg m^{-3})
3480	135.75	667.17	3.0790	– 1.5086	298.95	1.0438	– 0.9579	5566.89
3600	128.71	645.43	3.0955	– 1.5622	291.56	1.0543	– 0.9883	5507.52
3630	126.97	640.04	3.0997	– 1.5733	289.72	1.0569	– 0.9928	5492.63
3800	117.35	610.11	3.1246	– 1.6596	279.48	1.0726	– 1.0472	5408.72
4000	106.39	575.69	3.1563	– 1.7686	267.62	1.0926	– 1.1160	5309.63
4200	95.76	541.96	3.1911	– 1.8922	255.89	1.1146	– 1.1940	5209.55
4400	85.43	508.80	3.2297	– 2.0341	244.25	1.1389	– 1.1238	5108.10
4600	75.36	476.06	3.2730	– 2.1992	232.65	1.1663	– 1.3877	5004.69
4800	62.52	443.62	3.3221	– 2.3949	221.02	1.1972	– 1.5112	4898.69
5000	55.90	411.40	3.3786	– 2.6035	209.34	1.2329	– 1.6428	4789.64
5200	46.46	379.31	3.4446	– 2.9209	197.55	1.2745	– 1.8431	4676.95
5400	37.29	347.27	3.5231	– 3.2874	185.60	1.3241	– 2.0743	4559.93
5600	28.29	315.14	3.6187	– 3.7666	173.42	1.3844	– 2.3767	4437.64
5701	23.83	298.88	3.6756	– 4.0689	167.17	1.4203	– 2.5675	4373.62

Appendix G

A thermal model of the Earth

Table G.1 Thermal properties of the core

r (km)	T (K)	T_M (K)	γ	q	α (10^{-6} K^{-1})	C_P (J K^{-1} kg^{-1})	κ (W m^{-1}K^{-1})
0	5030	5330	1.387	0.1025	9.015	693	36
200	5029	5321	1.387	0.1027	9.033	693	36
400	5027	5294	1.388	0.1034	9.088	694	36
600	5023	5250	1.388	0.1045	9.181	695	36
800	5017	5188	1.389	0.1060	9.314	697	36
1000	5010	5107	1.390	0.1081	9.490	700	36
1200	5001	5012	1.391	0.1107	9.713	703	36
1221.5	5000	5000	1.391	0.1110	9.740	703	36
1221.5	5000	5000	1.390	0.1294	10.314	794	29.3
1400	4946	4890	1.391	0.1327	10.525	794	29.3
1600	4877	4772	1.393	0.1371	10.805	796	29.2
1800	4799	4629	1.395	0.1423	11.135	797	29.1
2000	4711	4545	1.398	0.1485	11.525	799	29.1
2200	4614	4452	1.401	0.1558	11.985	800	29.0
2400	4507	4261	1.405	0.1645	12.528	802	28.9
2600	4390	4057	1.409	0.1748	13.172	804	28.8
2800	4263	3840	1.415	0.1872	13.940	806	28.7
3000	4123	3609	1.421	0.2022	14.865	808	28.6
3200	3972	3367	1.429	0.2205	15.991	811	28.5
3400	3808	3112	1.438	0.2432	17.386	814	28.4
3480	3739	3007	1.443	0.2539	18.040	815	28.3

Table G.2 Thermal properties of the lower mantle

r (km)	T (K)	γ	q	α (10^{-6}K^{-1})	δ_S	ε	C_P (J K^{-1}kg^{-1})
3480	3739	1.1412	0.2770	11.290	1.2156	5.3816	1203
3600	2838	1.1447	0.2894	11.590	1.2405	5.4189	1191
3630	2740	1.1454	0.2926	11.667	1.2469	5.4289	1190
3800	2668	1.1515	0.3117	12.123	1.2848	5.4771	1191
4000	2596	1.1591	0.3365	12.708	1.3337	5.5170	1192
4200	2525	1.1676	0.3644	13.357	1.3879	5.5547	1193
4400	2452	1.1757	0.3919	14.149	1.4415	5.5633	1195
4600	2379	1.1881	0.4324	14.904	1.5172	5.6315	1196
4800	2302	1.2008	0.4748	15.848	1.5961	5.6730	1198
5000	2227	1.2154	0.5245	16.959	1.6872	5.7057	1201
5200	2144	1.2335	0.5860	18.255	1.7975	5.7556	1203
5400	2060	1.2548	0.6602	19.833	1.9285	5.8001	1206
5600	1974	1.2815	0.7549	21.801	2.0922	5.8465	1209
5701	1931	1.2972	0.8136	23.007	2.1921	5.8190	1214

Table G.3 Thermal properties of the continental upper mantle

r (km)	T (K)	γ	α (10^{-6}K^{-1})	C_P (J K^{-1}kg^{-1})
5701	2010	1.10	20.6	1200
5771	1985	1.11	21.3	1202
5871	1948	1.13	24.1	1209
5971	1907	1.15	27.4	1217
5971	1853	1.02	24.9	1202
6061	1817	1.04	26.9	1206
6151	1780	1.06	28.8	1210
6151	1719	0.96	30.2	1205
6256	1282	1.04	31.9	1197
6332	880	1.13	33.5	1186
6332	880	1.07	53.0	1208
6371	300	1.15	40.0	850

Appendix H

Radioactive isotopes

Table H.1 Naturally occurring, long-lived isotopes

Isotope	Percentage of element	Decay mechanism	Decay constant (year^{-1})	Half life (years)	Final daughter product
^{40}K	0.01167	85.5%β	5.544×10^{-10}	1.250×10^{9}	^{40}Ca
		10.5%K			^{40}Ar
		0.001%β^{+}			^{40}Ar
^{50}V	0.25	β	1.6×10^{-16}	6×10^{15}	^{50}Cr
		K			^{50}Ti
^{87}Rb	27.8346	β	1.42×10^{-11}	4.88×10^{10}	^{87}Sr
^{115}In	95.77	β	1.4×10^{-15}	5×10^{14}	^{115}Sn
^{123}Te	0.87	K	5.8×10^{-14}	1.2×10^{13}	^{123}Sb
^{138}La	0.089	β	6.3×10^{-12}	1.1×10^{11}	^{138}Ce
		K			^{138}Ba
^{142}Ce	11.05	α	1.4×10^{-16}	5×10^{15}	^{138}Ba
^{144}Nd	23.87	α	2.9×10^{-16}	2.4×10^{15}	^{140}Ce
^{146}Sm	~0.03	α	6.93×10^{-9}	1.0×10^{8}	^{142}Nd
^{147}Sm	15.07	α	6.54×10^{-12}	1.06×10^{11}	^{143}Nd
^{148}Sm	11.27	α	5.8×10^{-14}	1.2×10^{13}	^{144}Nd => ^{140}Ce
^{149}Sm	13.84	α	1.7×10^{-15}	4×10^{14}	^{145}Nd
^{152}Gd	0.20	α	6.3×10^{-15}	1.1×10^{14}	^{148}Sm => ^{144}Nd
^{156}Dy	0.0524	α	3.5×10^{-15}	2×10^{14}	^{152}Gd
^{176}Lu	2.60	β	1.87×10^{-11}	3.71×10^{10}	^{176}Hf
^{174}Hf	0.163	α	3.5×10^{-16}	2×10^{15}	^{170}Yb
^{187}Re	63.93	β	1.5×10^{-11}	4.6×10^{10}	^{187}Os
^{190}Pt	0.0127	α	1.16×10^{-12}	6×10^{11}	^{186}Os
^{204}Pb	1.364	α	4.95×10^{-18}	1.4×10^{17}	^{200}Hg
^{232}Th	100	$6\alpha + 4\beta$	4.9475×10^{-11}	1.4010×10^{10}	^{208}Pb
^{235}U	0.7201	$7\alpha + 4\beta$	9.8485×10^{-10}	7.0381×10^{8}	^{207}Pb
^{238}U	99.2743	$6\alpha + 4\beta$	1.55125×10^{-10}	4.4683×10^{9}	^{206}Pb
		5.4×10^{-5}% fission			

Note: ^{235}U and ^{238}U abundances do not add to 100% because ^{234}U occurs in the decay series of ^{238}U. A short-lived thorium isotope, ^{230}Th, also occurs in the decay series of ^{238}U. Intermediate daughters in the U and Th decay series are not included in this table. A list of the elements appears in Table A.3.

Table H.2 Short-lived isotopes, produced in the upper atmosphere by cosmic rays, arriving on the Earth with interplanetary dust or, in the case of ^{234}U, produced by radioactive decay, that are useful as tracers of atmospheric, oceanic or sedimentary processes

Isotope	Half-life	Decay mechanism	Decay product
neutron	10.6 minutes	β	^{1}H
^{3}H	12.26 years	β	^{3}He
^{7}Be	53.3 days	K	^{7}Li
^{10}Be	1.5×10^{6}	β	^{10}B
^{14}C	5730 years	β	^{14}N
^{22}Na	2.60 years	β^{+}	^{22}Ne
^{32}Si	160 years	β	^{32}S via ^{32}P
^{32}P	14.3 days	β	^{32}S
^{33}P	25 days	β	^{33}S
^{35}S	87 days	β	^{35}Cl
^{36}Cl	3.01×10^{5} years	β	^{36}Ar
^{37}Ar	35 days	K	^{37}Cl
^{39}Ar	270 years	β	^{39}K
^{53}Mn	3.8×10^{6} years	K	^{53}Cr
^{234}U	2.47×10^{5} years	α	^{230}Th => ^{206}Pb (half-life 75 000 years)

Table H.3 Extinct isotopes with decay products that are identifiable in meteorites or provide isotopic clues to early Solar System processes

Isotope	Decay mechanism	Half-life (years)	Decay constant (years^{-1})	Decay product
^{22}Na	β^{+}	2.60	0.267	^{22}Ne
^{26}Al	85% β^{+} 15% K	7.2×10^{5}	9.7×10^{-7}	^{26}Mg
^{60}Fe	β	3×10^{5}	2×10^{-6}	^{60}Ni via ^{60}Co
^{107}Pd	β	6.5×10^{6}	1.07×10^{-7}	^{107}Ag
^{129}I	β	1.6×10^{7}	4.2×10^{-8}	^{129}Xe
^{146}Sm	α	1.0×10^{8}	6.9×10^{-9}	^{142}Nd
^{182}Hf	2β	9×10^{6}	7.7×10^{-8}	^{182}W via ^{182}Ta
^{236}U	α	8.3×10^{7}	2.9×10^{-8}	^{208}Pb via ^{232}Th
^{244}Pu	99.7%α 0.3% fission	8.3×10^{7}	8.5×10^{-9}	99.7% ^{208}Pb (via ^{232}Th) 0.3% fission products

Appendix I

A geologic time scale

Table I.1 A geological time scale, 2004 (see Gradstein *et al.*, 2005). Numbers are dates of commencements of geological periods (millions of years ago)

Phanerozoic	Cenozoic	Quaternary	Holocene	0.01
			Pleistocene	1.81
		Tertiary	Pliocene	5.33
			Miocene	23.03
			Oligocene	33.9
			Eocene	55.5
			Paleocene	65.5
	Mesozoic	Cretaceous		145.5
		Jurassic		199.6
		Triassic		251
	Paleozoic	Permian		299
		Carboniferous		359
		Devonian		416
		Silurian		444
		Ordovician		488
		Cambrian		542
Proterozoic				2500
Archean				4000
Priscoan (Hadean)				

Appendix J

Problems

1.1 Calculate the moments of inertia of (a) a spherical shell and (b) a uniform sphere of mass M and radius R. (c) Consider a sphere within which density is inversely proportional to radius (giving gravity independent of radius, as in Problem 17.1, and a reasonable approximation for the Earth's mantle). What is the moment of inertia coefficient, I/MR^2?

1.2 (a) Consider a planet with a uniform mantle and a uniform core of half the total radius R, with density f times the mantle density. What values of f would be required to give moments of inertia $0.3307MR^2$, $0.365MR^2$ and $0.391MR^2$, corresponding to the Earth, Mars and Moon respectively?

(b) Consider instead that the ratio of core to mantle densities is 3.0, both being uniform. What fractions of the total radius must the core have in the three cases?

1.3 Allen (1973) gives the following model for the density of the Sun as a function of radius r. The outer radius is $r_S = 6.9 \times 10^8$ m and the total mass is $M_S = 1.989 \times 10^{30}$ kg. Use the values to obtain a rough estimate of the moment of inertia of the Sun. (Allen's estimate from finer details is 5.7×10^{46} kg m^2.) What fraction is this of the moment of inertia of a uniform sphere of the same mass and radius?

r/r_S	ρ(kg m^{-3})	r/r_S	ρ(kg m^{-3})
0	160 000	0.6	350
0.04	141 000	0.7	80
0.1	89 000	0.8	18
0.2	41 000	0.9	2
0.3	13 300	0.95	0.4
0.4	3600	1.0	0
0.5	1000		

1.4 By virtue of the rotation of the Sun, radiation from opposite (approaching and receding) limbs is seen to be Doppler shifted (to blue and red respectively). Thus the radiation field carries away some of the solar angular momentum. What is the fractional rate of slowing of the Sun's rotation by this effect? (Use the solar moment of inertia from Problem 1.3 and other parameters from Table A.5, Appendix A.)

1.5 Consider a spherical black body at 1 AU from the Sun to be rotating and tumbling in such a way as to equalize the temperature over its whole surface. What is that temperature? (The solar constant is 1370 W m^{-2}.)

1.6 What would be the equilibrium temperature of the subsolar point on the body in Problem 1.5 if it presented a constant face to the Sun? (Neglect conduction within the body.)

1.7 What is the equivalent black-body temperature for the surface of the Sun (radius 6.96×10^8 m)?

1.8 How does the equilibrium temperature of a planet depend on the radius of its orbit?

1.9 Consider a space-craft approaching the Solar System in a search for planets. As it gets closer it begins to observe the planets in an order according to the total amount of sunlight reflected. Assuming that, in the wavelength range used, all the albedos (reflection coefficients) are equal, what is this order? (Planetary parameters are listed in Table 1.1.)

1.10 Suppose that collisions in the asteroidal belt are continuously producing interplanetary dust with a distribution of sizes such that the number of particles dN within any mass range m to $(m + \mathrm{d}m)$ is given by $\mathrm{d}N \propto m^{-n}\,\mathrm{d}m$, where $n > 1$. Once produced, the dust spirals towards the Sun by the Poynting–Robertson effect (Eq. (1.21)). What is the distribution of particle masses intercepted by the Earth? Assume the mass intercepted to be only a small fraction of the total and to be determined by residence time near to the Earth's orbit and that all particles have equal densities. (Note the evidence, discussed in Sections 1.8 and 1.9, that the flux of dust originating in the asteroidal belt is not steady.)

1.11 What is the average speed of arrival on the Earth of meteoroids falling freely from outside the Solar System? How does it compare with the escape speed from the Earth?

2.1 Assuming meteorites to be of two kinds, irons that are 100% metal and chondrites that average 10% metal, and that the total metal content of the meteorite bodies is the same fraction of their total mass as the core is for the total of the Earth, what fraction of the total mass of meteorites is in the irons? (Masses of the Earth and the core are given in Appendix A, Table A.4.)

2.2 The uncompressed density of the planet Mercury is estimated to be 5280 kg m^{-3}. If it consists of a core and mantle with compositions similar to those of the Earth, what is the ratio of core radius to total radius?

2.3 The elements in Table 2.1 are listed in the order of their abundances in the solar atmosphere. Below are the nuclear mass excesses for their most abundant and next most abundant isotopes. The reference standard (zero) is the isotope ^{12}C and the other values are departures of the mass per nucleon from the ^{12}C mass, in parts per million. The atomic weights are given in parentheses. Is there a direct correlation between abundances and nuclear mass excesses? What other factors need to be taken into account? Note: see the proton and neutron masses in Table A.1 (Appendix A).

H	(1)7825	(2)3913
He	(4)651	(3)5343
O	(16) −318	(18) −47
C	(12)0	(13)258
Ne	(20) −378	(22) −392
Fe	(56) −1162	(54) −1118
N	(14)220	(15)7
Si	(28) −824	(29) −811
Mg	(24) −623	(26) −670
S	(32) −873	(34) −945
Ar	(36) −902	(40) −940
Ni	(58) −1115	(60) −1154
Ca	(40) −935	(44) −1012
Al	(27) −684	–
Na	(23) −445	–

3.1 (a) Calculate the gravitational energy released by the collapse of a mass M of material, initially dispersed to infinite separation, to a uniform sphere of radius R.

(b) Repeat the calculation for the Sun, assuming the density distribution in Problem 1.3.

3.2 Derive Eqs. 1.8 and 1.9.

3.3 Show that for atoms of a radioactive species with decay constant λ, the mean life is λ^{-1}.

3.4 Consider a rock that yields the following isochron data:

$\mathrm{d}^{40}\mathrm{Ar}/\mathrm{d}^{40}\mathrm{K} = 0.098 \pm 0.005$,

$\mathrm{d}(^{87}\mathrm{Sr}/^{86}\mathrm{Sr})/\mathrm{d}(^{87}\mathrm{Rb}/^{86}\mathrm{Sr}) = 0.0215 \pm 0.0007$,

$\mathrm{d}(^{207}\mathrm{Pb}/^{204}\mathrm{Pb})/\mathrm{d}(^{206}\mathrm{Pb}/^{204}\mathrm{Pb}) = 0.090 \pm 0.005$.

(a) Estimate the age and uncertainty, from each of these isochrons.

(b) Suggest reasons for the discrepancies between the alternative estimates and a probable history.

3.5 The decay constants for the uranium isotopes are known with greater accuracy than the decay constant for ^{87}Rb. Use the

Rb–Sr isochron for meteorites (Eq. (4.4)) with the meteorite age, 4.54×10^9 years, from the Pb–Pb isochron (Eq. (4.3)) to calculate the decay constant, λ, for ^{87}Rb. Assuming that the uncertainty in this calculation is due entirely to scatter of values in the Pb–Pb isochron, what is the uncertainty in λ for ^{87}Rb?

4.1 If the processes of nuclear synthesis produced ^{127}I and ^{129}I in equal atomic abundances and proceeded at a uniform rate for a time τ, and did not operate either before or after this, and if accumulation of ^{129}Xe in a particular meteorite commenced at time t after the cessation of nuclear synthesis, obtain the expression relating the ratio (^{129}Xe/^{127}I) in the meteorite to τ, t, and the decay constant λ for ^{129}I. Show that Eq. (4.6) is the special case for $\tau \ll \lambda^{-1}$.

4.2 If the heavy elements were formed 4.67×10^9 years ago (i.e. about 10^8 years before solidification of the meteorites), what was the ratio of ^{235}U/^{238}U at that time? (Note: the estimate includes all heavier species that decayed to these isotopes. The original ^{235}U abundance itself was probably very small because, in a neutron-rich environment, ^{235}U undergoes rapid neutron-fission.)

4.3 Two chondrules, A and B, from the same meteorite have ratios of excess ^{129}Xe to ^{127}I differing by a factor two, the value for A being larger. If this is interpreted as a difference in formation dates, which chondrule formed first and by how long? (See Table H.3 for ^{129}I decay.)

4.4 Measurements on a freshly fallen meteorite give an isotopic ratio $^3\mathrm{He}/^3\mathrm{H} = 4.52 \times 10^5$. Noting that ^{3}H decays to ^{3}He and assuming that the two isotopes were produced in equal abundances by cosmic ray bombardment, what was the duration of the cosmic ray exposure of the Meteorite?

5.1 Use the constants in Table H.1 to derive Eqs. (5.2) and (5.3).

5.2 Consider a model for the growth of ^{40}Ar in the Earth and the atmosphere, according to which the Earth starts with no argon t years ago, but accumulates it by decay of ^{40}K, of which there is now a mass K remaining. Argon is lost to the atmosphere at a rate proportional to its concentration in the Earth, that is the rate is Λ times the total argon in the Earth.

(a) How much argon is there in the Earth now, in terms of K, Λ, and the decay constant, λ, for ^{40}K?

(b) What is the argon content of the atmosphere?

(c) What is the ratio of (b) to (a)?

(d) If $t = 4.5 \times 10^9$ years and the argon contents of the Earth and atmosphere are equal, what is the value of Λ?

(To solve a differential equation of the form $\mathrm{d}y/\mathrm{d}x = Ae^{ax} + By$, put $y = ue^{ax}$.)

5.3 What is the gravitational energy release resulting from settling of a core of density $2\bar{\rho}$ and radius $0.55R$ from an initially uniform Earth of density $\bar{\rho}$ and radius R, assuming as an approximation that the density of each element of material is unaffected? Use values for mass and radius of the Earth. What average temperature rise would this energy release cause? Assume an effective heat capacity of $9.93 \times 10^{27}\ \mathrm{J\,K^{-1}}$. The result of a more rigorous calculation is given as an entry in Table 21.1.

6.1 If the whole of the angular momentum of the Solar System, $3.15 \times 10^{43}\ \mathrm{kg\,m^2\,s^{-1}}$, were put into the Sun, without changing its size or density profile, would it be rotationally stable? This question can be answered at two levels. The simpler one is to assume that the Sun remains spherical and compare the gravitational attraction with the centrifugal ‘force’ at the equator. A better approximation is to allow the Sun to deform to an equilibrium oblate ellipsoid and use the theory of Sections 6.3 or 8.5 to calculate the flattening. A value of flattening, f, less than unity ($c/a > 0$) indicates stability. Assume the solar moment of inertia given in Problem 1.3.

6.2 Consider the Moon to be momentarily directly between the Sun and Earth (during a solar eclipse) and calculate the ratio of its gravitational attractions to the Sun and

Earth. Since the solar gravity is stronger, explain how the Moon remains in orbit about the Earth instead of going into an independent orbit about the Sun.

6.3 Consider a fluid planet, of uniform density ρ, rotating with angular speed ω, giving a *slight* consequent flattening, f. Show that $f \propto \rho^{-1}$. (Note: this is the reason why the core is less flattened than the Earth as a whole.) Show that your result is consistent with Eq. (6.39).

6.4 Show that, for a planet of uniform density, or one in which the surfaces of equal density are homologous ellipsoids (all having the same ellipticity), the dynamical ellipticity, H (Eq. (7.2)), is related to the surface flattening, f, by

$$H = f\left(1 - \frac{1}{2}f\right).$$

6.5 (a) For Mars $\omega = 7.0882 \times 10^{-5}\,\mathrm{rad\,s^{-1}}$ and $J_2 = 1.825 \times 10^{-3}$. Assuming hydrostatic equilibrium, calculate f and C/Ma^2. (For mass and radius see Table 1.1, with values for the Earth in Table A.4).

6.6 Show that for a spherical planet of mean density $\bar{\rho}$, equatorial gravity vanishes at rotational speed given by

$$\omega_{\mathrm{crit}} = \left(\frac{4}{3}\pi G\bar{\rho}\right)^{1/2},$$

and that if the density is uniform the corresponding angular momentum is

$$L_{\mathrm{crit}} = 0.32 G^{1/2} M^{5/3} \rho^{-1/6},$$

where M is planetary mass. It appears that many planetary and asteroidal angular momenta roughly follow a law of this form, but with a numerical coefficient of about 0.07, that is 20% of the critical value. What values of this coefficient apply to the Earth and to Jupiter (for which you will need to estimate the moment of inertia). What reason is there for the Earth to have a low value?

6.7 (a) If the following gravity measurements were made at noon each day on a ship sailing due north at constant speed, at what time on which day did it cross the equator and what was its speed?

Day	g	Day	g
1	9.802 22	9	9.780 50
2	9.798 05	10	9.780 33
3	9.794 15	11	9.780 83
4	9.790 59	12	9.781 97
5	9.787 47	13	9.783 74
6	9.784 84	14	9.786 09
7	9.782 78	15	9.788 97
8	9.781 32	16	9.792 32

(b) If at latitude φ the ship turned due East, what was the resulting change in gravity measured on the moving ship?

7.1 The equator of Mars is inclined at 24° to the orbital plane. Using parameters from Problem 6.5, what is the rate of precession due to the solar torque?

7.2 By making guesses for the total mass and mean speed of traffic and the average separation of opposing traffic lanes, estimate the effect on the Earth's rotation of the diurnal development of traffic in North America.

7.3 (a) Consider two concentric spheres that are constrained to rotate about axes that are inclined to one another by an angle ϕ. If a thin layer of viscous fluid between them exerts a linear frictional drag on the differential motion and the outer sphere is maintained at angular speed ω_0, what is the rotational speed of the inner one in equilibrium with it?

(b) If the fluid layer has radius r, thickness $d \ll r$ and viscosity η, what is the power dissipation?

(c) If there is a 'viscous' coupling coefficient between the core and mantle given by Eq. (7.33) and precession maintains an angular difference of 6×10^{-6} rad between the core and mantle axes, what is the consequent dissipation?

(d) What effect does the dissipation in (c) have on the Earth's rotation? Is this plausible? If plausible and a layer with core viscosity $\sim 10^{-2}$ Pa s is responsible, what is its thickness?

7.4 Using Eq.(7.22), estimate the elastic strain energy associated with the Chandler wobble, as a fraction of the total wobble energy.

7.5 Assuming equilibrium flattening of the Earth, proportional to ω^2, how does the period of the Chandler wobble depend on ω? What rate of rotation is required to make the wobble period 1 year? When might this have occurred?

7.6 Consider a mass m of water to be impounded in a reservoir at latitude ϕ, effectively withdrawing it from the ocean.

(a) With the simple assumption that the axis of rotation is unchanged, what value of ϕ would give no change in LOD?

(b) When considered more closely, the centre of mass of the Earth and the axis of maximum moment of inertia are changed. By how much?

(c) If the reservoir is filled in less than half a wobble period (7 months), then its filling would contribute to excitation of the Chandler wobble. What mass of water and latitude would be required to make this significant?

8.1 Substituting Eq. (8.2) in Eq. (8.1) we find a $\cos\psi$ term in W. Why does an asymmetrical term of this form not appear in Eq. (8.9)?

8.2 Show that for a fluid Earth of uniform density, $k_2 = 3/2$.

8.3 What is the approximate amplitude of the tide raised in the planet Mercury by the Sun? How does this compare with the probable elastic limit?

8.4 Suppose that the primeval Earth had a rotation period half of the present value and that subsequent slowing has been due entirely to angular momentum exchange with the Moon, all motions being coplanar.

(a) If the Moon was formed in a circular orbit, what was the orbital radius?

(b) If the Moon was captured from a parabolic orbit (as has sometimes been suggested), what was the distance of its closest approach?

8.5 Assuming conservation of angular momentum in the Earth–Moon system, what will be the angular velocities of the Earth's rotation and the Moon's orbital motion and the distance of the Moon when the tide in a 5 km deep equatorial channel is resonant? (In water of depth h, the speed of a wave of wavelength $\lambda \gg h$ is $\sqrt{gh}$.)

8.6 If the Moon were rotating at one revolution per day at its present distance from the Earth, with a tidal phase lag $\delta = 0.2^\circ$ in Eq. (8.20) and following equations, and $k_2 = 0.3$, how long would it take for tidal friction to stop the rotation (relative to the Earth)?

8.7 The tidal potential Love number for the Earth as a whole, as observed by satellites, is $k_2 = 0.245$, with a phase lag, $\delta = 2.9^\circ$, of the tidal bulge. As noted in Section 8.2, for the solid Earth alone the Love number would be $k_S = 0.298$. The difference represents an inverted tide of the oceans, with negative k_O. Marine tides are complicated, but we can regard k_O as a vector representing their total effect. Tidal dissipation is dominated by the oceans, so we can consider k_S as a vector aligned with the Moon (or Sun). What are the magnitude and phase of k_O, as the vector difference between k_2 and k_S? What amplitude of the marine tide does this correspond to?

9.1 (a) If the Greenland ice sheet (at 75°N) decreases by the melting of 1000 km^3 of ice annually, distributing melt water uniformly over the Earth with no other adjustment, estimate

(i) the angular migration of the pole,

(ii) the associated slowing of the Earth's rotation.

(b) What influence would isostatic rebound have on the conclusion to part (a)?

9.2 (a) Calculate the gravity due to a thin disc (thickness Δz) of density ρ and radius R at a point on its axis at distance h from it.

(b) What is the simple approximation for $R \gg h$, but $R < \infty$?

(c) Use the result in (a) or (b) to obtain the gravity due to an infinite sheet. Note that, since the result is independent of h, there is no restriction on Δz in this case.

(d) Why cannot the result in (c) be obtained by considering a spherical shell and allowing its radius to become infinite?

9.3 In the 1850s G. B. Airy, whose explanation of isostasy is illustrated by Fig. 9.6(b), reported an experiment to determine the value of the Newtonian gravitational constant, G, from the variation of gravity, g, with depth in a mine and crustal density, ρ. The results were reported as a measurement of the mean density of the Earth, $\overline{\rho}$.

(a) Why is the determination of $\overline{\rho}$ equivalent to the determination of G?

(b) What is the ratio $\rho/\overline{\rho}$ that would give zero gravity gradient in a mine?

(c) Crustal density is typically half of the mean Earth density. What is the typical crustal gravity gradient?

(d) What is the gravity gradient in the ocean?

(e) By how much does the free air gradient differ from the hypothetical 'free vacuum gradient', with no atmosphere?

10.1 Obtain the expression for Poisson's ratio in terms of the ratio of P- and S-wave speeds. (See Appendix D for relationships between moduli.)

10.2 Consider a laminated medium composed of alternate layers with equal thicknesses of material with P-wave speeds V_1 and V_2 for bulk material. The values of density and Poisson's ratio are the same for both materials. Show that for P-waves with wavelengths that are large compared with the layer thicknesses (but small compared with the overall dimensions of the medium) the velocities perpendicular and parallel to the layers are

$$V_{\text{perpendicular}} = \frac{\sqrt{2}V_1V_2}{(V_1^2+V_2^2)^{1/2}},$$

$$V_{\text{parallel}} = \left\{\frac{V_1^2+V_2^2}{2}\left[1-\frac{\nu^2}{(1-\nu)^2}\frac{(V_1^2-V_2^2)^2}{(V_1^2+V_2^2)^2}\right]\right\}^{1/2}.$$

What values of V_1/V_2 are required for velocity anisotropies of 1%, 5%?

(Hint: calculate the elastic moduli for both cases, using Eqs. (10.4) for each layer. The expression for Young's modulus, E, in terms of χ and μ is given in Appendix D.)

10.3 Consider two seismic stations, one close to an earthquake and the other on the opposite side of the Earth. In a particular frequency range the first station sees a white P-wave spectrum and the spectrum from the remote one gives d ln (amplitude)/ d(frequency) = −6. Assuming Q to be independent of frequency,

(a) what is the effective Q_P for the trans-Earth path?

(b) what is the average Q_P for the mantle if the core has infinite Q?

(c) what value of Q_S for the mantle would be expected?

(d) what value of d ln(amplitude)/d(frequency) would the S-wave spectrum at the remote station give, if the S-wave spectrum at the near station is white?

10.4 Assuming elliptical anelastic hysteresis loops (Fig. 10.3), what resolution in strain recording is required to measure to 10% a rock Q of 200, using a strain cycle of amplitude 10^{-6}.

11.1 Solve for the stresses and displacements for a 10 × 10 × 10 km elastic block in the gravitational field of the Earth, with zero friction at its base. What is the maximum displacement if the elastic moduli are given by $\mu = \lambda = 3 \times 10^{10}$ Pa?

11.2 Consider a granular material which has a cohesion S_0 and coefficient of friction μ_f, subject to a body force that gives rise to a vertical stress $\sigma_{zz} = \rho gz$. Find, from the Coulomb shear stress, the angle of repose above which sliding occurs.

11.3 An approximate solution to the stresses from a pressurized spherical magma chamber of radius R and internal pressure P has radial stress given by $\sigma_{rr} = \dfrac{AP}{r^3}$ and a tangential stress given by $\sigma_{\theta\theta} = \sigma_{\phi\phi} = -\dfrac{1}{2}\dfrac{AP_0}{r^3}$.

(a) Find the coefficient A, and give an expression for the displacement field.

(b) Now consider the case of a spherical cavity with radius R and zero internal pressure in an infinite medium. Let the pressure at infinity be a uniform pressure. Show that radial and tangential stresses are

$$\sigma_{rr} = P\left(1 - \frac{R^3}{r^3}\right),$$

$$\sigma_{\theta\theta} = P\left(1 + \frac{R^3}{2r^3}\right),$$

11.4 Suppose that a hydrofracture experiment is carried out at various depths in a bore hole, and that above a depth of 1 km the fractures are horizontal, but below 1 km they are all vertical. Let the maximum stress, σ_1, be horizontal and equal to 100 MPa. Estimate the value of all the principal stresses at a depth of 2 km.

11.5 Consider a hydrofracture experiment at 1 km depth in a medium with density $\rho = 3000\,\mathrm{kg\,m^{-3}}$. A crack extends north and south of the borehole. The break-down pressure $P_b = 40$ MPa. The shut-in or closure pressure is $P_c = 25$ MPa. The vertical principal stress is determined by gravity (assume $g = 10\,\mathrm{m\,s^{-2}}$).

(a) Calculate the principal stresses and their directions.
(b) What type of earthquake would be expected in this region? Explain.
(c) If earthquakes were to occur in this region and the coefficient of friction is 0.6, what would be the expected geometry of the fault planes?.

11.6 An approximate solution to the stresses from a pressurized cylindrical magma chamber has radial stress given by $\sigma_{rr} = \frac{AP_0}{r^2}$ and a tangential stress given by

$$\sigma_{\theta\theta} = -\frac{AP_0}{r^2}, \sigma_{\phi\phi} = 0.$$

(a) Show that these equations satisfy the 2D equilibrium equations.
(b) If P_0 is the pressure at $r = R$, find the coefficient A.
(c) Derive an expression for the displacement field.

11.7 The stresses caused by a line load (between $x = -\infty$ and $x = \infty$) of forces applied in the horizontal direction at the surface of an elastic half space, $z > 0$, are

$$\sigma_{zz} = \frac{2Xz^2y}{\pi r^4}; \sigma_{yy} = \frac{2Xy^3}{\pi r^4}; \sigma_{zy} = \frac{2Xzy^2}{\pi r^4}. \qquad \text{(J.1)}$$

(a) Show that these equations satisfy the equilibrium equations.
(b) Confirm that they satisfy the free surface boundary conditions (zero normal and tangential stresses) except along the line of application of the force.
(c) If the force has strength F per unit length on the x axis, by integrating horizontal traction about a cylindrical region surrounding the line of application, find the value of X in terms of this force.

11.8 Stresses generated by a normal line load of strength X per unit length at the surface of an elastic half space are given by

$$\sigma_{zz} = \frac{2Xz^3}{\pi r^4}; \sigma_{yy} = \frac{2Xzy^2}{\pi r^4}; \sigma_{xy} = \frac{2Xz^2y}{\pi r^4}.$$

Show that this is equivalent to a horizontal line $(-\infty < y < \infty)$ of vertically directed forces (z – direction), strength F per unit length, by integrating stress about a cylindrical region surrounding the line of application and find the value of X in terms of F.

11.9 The stresses in an elastic half space caused by a horizontal line load at the surface are given in Problem 11.7. Use these to consider the stresses arising from a long strip of material of thickness h and width $2W \gg h$ that is welded to a half space, $z > 0$, with the same elasticity, and is heated, causing it to expand thermally. Changes of the dimension in the direction of its length x are prevented, and the surface is free. The change in the width of the strip (in the y direction) is constrained by the half space. Show that along the mid-line of the strip the y strain is $(4/\pi)(h/W)e$, where e is the strain that would occur without the constraint of the half space..

(a) First use Eqs. (10.24) to show that the elastic modulus describing the plane strain situation (with zero strain maintained in the x direction) is $E/(1 - \nu^2)$.
(b) The problem can be solved by first considering the unconstrained strip to be compressed by line forces along its

edges of magnitude to reduce the strain to the constrained value, that is by causing strain $e[1-4\pi(h/W)]$. Then apply opposite line forces of the same magnitude a distance $2W$ apart in the surface of the half space to produce strain $e(4\pi)(h/W)$ in the half space at the midpoint line, so that when the strip and the half space are welded the forces cancel and the strain $e(4\pi)(h/W)$ remains.

12.1 DeMets *et al.* (1990) give the motion of the Indian plate relative to the Pacific plate as 1.1539×10^{-6} degrees per year about a pole at (60.494N 30.403W). Show that, with the motion of the Pacific plate relative to the Hawaii hot spot given as the first entry in Table 12.1, the motion of the Indian plate given in the table follows.

12.2 If the ocean floor heat flux, discounting hot spot heat attributable to core cooling, 27 tW, is explained by cooling, to a depth of 100 km, of lithosphere that is replaced at a rate of 3.4 km^2/year, what is its average temperature change? Is this reasonable, or should the estimated depth of cooling be adjusted?

12.3 Bird (2003) estimated the total linear length of spreading centres to be 67 000 km and zones of convergence (subduction) to be 51 000 km. if we take the sum of them to be the total length of plate boundaries, what is the mean plate size, that is the 'reach' or distance between source and sink? Is this consistent with the estimates of mean plate speed in Section 13.2 and average plate lifetime from Section 20.2 (90 million years)? Such average numbers can be misleading. There is a range of plate sizes and speeds and the relationship between averages depends on how size and speed, or lifetime and speed, are correlated. Suppose the speeds to be uniformly distributed over the range 0 to $\nu_{\max}$ with an average $\bar{\nu}=\nu_{\max}/2$, and approximate this with a distribution in which half of the ridge length produces crust spreading at $\bar{\nu}/2$ and the other half produces crust spreading at $3\bar{\nu}/2$. Suggest how the lifetimes of the plates may be correlated with their speeds and show how the correlation influences the relationship between mean values.

12.4 Consider a continental block of thickness 40 km and diameter 4000 km, with a surface elevation 840 m above sea level, isostatically balanced with ocean floor at an elevation of −4500 m, to drift from Antarctica to the equator. Estimate the change in the Earth's rotation rate and the shift of the pole of rotation with respect to all other features, which remain fixed in relative positions, under two alternative conditions, (a) there is no change in ellipticity, (b) the equatorial bulge readjusts to equilibrium.

12.5 Some geological reports indicate that, 100 million years ago, sea level was ~200 m higher than a present. We assume this to be a global effect and not a local effect of land movement. It cannot be accounted for by complete ice cap melting (which would cause a rise of about 80 m) combined with any plausible warming and thermal expansion of the oceans. Consider the changes in ocean floor topography that could have been responsible. Isostatic balance of continents and ocean basins ensures that there was no net transfer of mass between such large areas and therefore that sea level variations would have been caused by variations in the thermal structure of the lithosphere. The diffusive cooling model for the oceanic lithosphere (Section 20.2) gives the total cooling, and therefore shrinkage, at age t as proportional to $t^{1/2}$. This causes ocean deepening, relative to the ridge crests, $z=z^*(t/\tau)^{1/2}$, where $z^*=3000$ m is the value at age $\tau^*=90$ million years, the present average duration (lifetime) of the oceanic lithosphere. Flooding of the continents would be caused by a decrease in the average value of z, and therefore $t^{1/2}$, requiring the ocean floors to be hotter and younger, on average, than at present. This could have arisen in several ways. One possibility is a plate redistribution such that average plate ages are now more nearly equal than 100 million years ago, but with no change to the overall average, effectively replacing with middle-aged lithosphere areas that were previously both younger and older. Consider two simpler alternatives, in the approximation that all plates have the same lifetime: (i) faster plate motion with no change in geometry, and (ii) more (smaller) plates, with correspondingly more ridges and subduction zones, and no change in plate speed. By how much must the lifetime decrease to give a 200 m rise in sea level? What is the implied change in heat flux? In Section 23.2, we allow for hydrothermal cooling by replacing $t^{1/2}$ in the above expressions with $t^{1/3}$. How much difference would this make?

13.1 Using Eq. (13.6) and Fig. 20.2, with the isostatic correction factor, 1.437, as applied to Eq. (20.11), estimate the absolute minimum age of oceanic lithosphere that may subduct.

13.2 Assuming that the rate of subduction of lithosphere is 3.36 km^2/year and that its thermal shrinkage averages 2.1 km, calculate the rate at which energy would be released by subduction to the base of the mantle of all the coolness that this represents. Approximations that suffice for this purpose are $g = 10\,\mathrm{m\,s^{-2}}$ (constant) and the variations in density and thermal expansion coefficient are such that the product, $\alpha\rho$, decreases linearly with depth, by a factor 2 over the whole mantle depth. Compare your result with the 7.7×10^{12} W of convective energy derived thermodynamically in Chapter 22. Your answer will be about twice this allowed maximum. What is the explanation?

13.3 Calculate the gravitational energy released by the separation of the crust from the mantle. This is energy of compositional convection in the same sense as the energy released by redistribution of light solute by the solidifying inner core (Section 22.6) and is 100% efficient in contributing to the mechanical energy of convection. How significant a contribution is it? Compare your answer with the relevant entry in Table 21.1.

13.4 Ignoring core heat, and assuming mantle convection to be driven only by its own internal radiogenic heat, over what depth at the base of the mantle would this heat be inadequate to drive convection? Assume the total mantle radiogenic heat to be 20×10^{12} W and uniform conductivity $5\,\mathrm{W\,m^{-1}K^{-1}}$. To calculate the temperature gradient, use Eq. (19.55) with thermal properties in Appendix G.

13.5 If the viscosity of plume material is (a) underestimated or (b) overestimated by a factor 100 in Section 13.3, what would be the corrected plume diameter for each case?

13.6 Derive Eq. (13.20), using Eq. (13.25).

13.7 The slope of the accretionary wedge that forms the high Andes (roughly latitudes 13°S to 27°S) is $\alpha = 4^\circ$, twice that of the low Andes, further north and south, for which $\alpha = 2^\circ$. It has been proposed that this difference arises from the cumulative effect of greater friction on the top of the down-going plate (Lamb and Davis, 2003) beneath the high Andes. Because of climate conditions the trench in front of the central Andes is almost devoid of sediment, whereas to the north and south the trench has thick sediments. Assume that the amount of sediment determines the frictional properties on the slab, mainly by water pressure both in the wedge and at its base, and that the effective friction beneath the high Andes is double that beneath the low Andes. Use Eq. (13.41) to analyse these differences.

(a) Assuming that for the low Andes $\lambda_w = 0.7$ and $\mu_w = \mu_b = 0.85$ (as used for Taiwan in Section 13.6) and that $\beta = 6^\circ$, what is the value of λ_b for the low Andes that gives rise to $\alpha = 2^\circ$?

(b) What values of λ_b and λ_w for the high Andes give $\alpha = 4^\circ$, along with the doubled friction on the base? Comment on whether the solution is reasonable, given the arguments above.

14.1 Equation (14.1) gives strain as a function of distance from the axis of a screw dislocation (of infinite length) in a uniform (infinite) medium. Using the fact that strain energy per unit volume is ½$\mu\varepsilon^2$, where μ is rigidity, calculate the strain energy per unit length of a screw dislocation. What is wrong with the assumptions in the problem that cause the answer to tend logarithmically to infinity? How can the question be posed more realistically?

14.2 Equation (14.34) defines surface-wave magnitude, M_S, and Eq. (14.36) gives an empirical relationship to earthquake energy. Using the fact that $(2\pi a/T)$ is the peak ground strain in a seismic wave and the square of this gives the wave energy density, relate the total energy to the wave energy density. What does the result imply about the variation with magnitude of the duration of the seismic wave train? Is this plausible? Note that the same conclusion does not follow if body-wave magnitudes m are used with the equation equivalent to (14.36):

$$\log_{10} E = 2.3\,m - 0.5.$$

14.3 Consider a train of P-waves of period T, amplitude a and total duration τ, observed at distance R from an earthquake. The local rock has density ρ and P-wave speed V_P. If these observations are representative of all directions from the earthquake, what is the total energy in the P-waves?

14.4 Use the general expression for the phase speed of an ocean wave, v (Eq. (14.53)), with the expression for group velocity, u (Eq. (16.49)), to obtain the relationship between u and v for water of arbitrary depth, h. Show that for long wavelengths ($\lambda \gg h$) $u = v$ and that for short wavelengths ($\lambda \ll h$) $u = v/2$.

14.5 Using the equation for displacements by a unit point force

$$G_{ik} = \frac{1}{4\pi\mu}\left[\frac{\delta_{ik}}{r} - \frac{1}{4(1-\nu)}\frac{\partial^2 r}{\partial x_i \partial x_k}\right],$$

(a) Show that the displacements are of the form

$$u_x = B\left(\frac{x^2}{r^3} + \frac{(\lambda + 3\mu)}{r(\lambda + \mu)}\right);\, u_y = \frac{Byx}{r^3};\; u_z = \frac{Bxz}{r^3};$$

and show that $B = \dfrac{1}{4\pi\mu}\dfrac{1}{4(1-\nu)}$ (for unit point force B has units of length^2).

(b) Obtain the static displacement field for a double couple source with slip in the x-direction across a plane oriented in the y-direction.

(c) Assume that a small earthquake has a displacement b on area S in a medium of modulus μ and show that the displacement field at any point in the medium is given by

$$u_x = \frac{\mu b S}{12\pi\mu}\frac{y}{(y^2 + c^2)^{3/2}} = \frac{bS}{12\pi}\frac{y}{(y^2 + c^2)^{3/2}}.$$

(d) Consider a magnitude 6 strike-slip earthquake on a square fault plane oriented east–west, with its top boundary at a depth of 4 km. Seismic records indicate a stress drop of 3 MPa. With the assumptions that the slip is uniform across the fault plane and that the rigidity of the medium is $\mu = 3 \times 10^{10}$ Pa, use the moment–magnitude relationship and the definition of moment to estimate the dimension of the fault plane and magnitude of slip.

(e) Calculate and plot the horizontal displacements on the surface at $x = 0$, $y = -5, -4, -3, -2, 0, 1, 2, 3, 4, 5$ km.

14.6 Integrate Eq. (14.27) for different time intervals to obtain the trapezoid shapes in Figure 14.11(c).

14.7 The Brune (1970) spectrum (Eq. (14.32)) of a far field seismic pulse corresponds to a seismic pulse of shape given by $P(t) = kt \exp(-at)$. Show that

(a) $k = u(0)\omega_0^2$,

(b) $a = \omega_0$,

(c) the moment is $M_0(t) = M_0[1 - (at + 1) \times \exp(-at)]$.

15.1 If a transcurrent (transform) fault, across which there is a shear stress of 10^7 Pa, creeps aseismically at an average rate of 5 cm/year, estimate the magnitude of the localized heat flow anomaly along it. Assume that in the uppermost layer the shear stress is reduced to the same value as the normal stress. (The reason for lack of a heat flow anomaly of the estimated magnitude is considered in Section 15.6. Measurements along the San Andreas fault in California drew attention to this problem.)

15.2 Can the absence of a heat flow anomaly, referred to in Problem 15.1, be explained by retention of heat (or energy in any other form) in the fault zone? Calculate the total heat generated by friction if the slip accumulates to 200 km and the heat does not escape? If it is applied to the melting of granite, latent heat $4.5 \times 10^5\,\mathrm{J\,kg^{-1}}$, estimate the thickness of rock on either side of the fault that would melt.

(b) use Eq. (15.41) to show that $f_0 = (\zeta V_R/l)/2\pi = 3.5 \times 0.9 \times V_S/(2\pi l)$, and that

$$M_0 = \Delta\sigma\left(\frac{3.15}{2\pi}\right)^3\left(\frac{V_S}{f_0}\right)^3;$$

(c) show that the radiation efficiency is 49.7%.

15.5 (a) Give an expression for corner frequency as a function of magnitude.

(b) Assuming that earthquakes have an average stress drop of 3 MPa in rocks of rigidity $\mu = 3 \times 10^{10}\mathrm{Pa}$, use formulae from Chapters 14 and 15 to complete the following table. Assume that faults are equi-dimensional and that rupture velocity is $3\,\mathrm{km\,s^{-1}}$.

Magnitude	Moment	*l* (Size)	*T*, rupture time	*b* (slip)	f_c (corner frequency)
4					
5					
6					
7					
8					

15.3 Let the time to failure for aftershocks be $t_{\text{failure}} = A\tau^{-n}$, where τ is stress and n is a large positive integer. Show that this gives rise to Omori's law (Eq. (15.26) or (15.27)) for aftershocks, and give the restrictions on the distribution of stress in the aftershock zone.

15.4 Calculate the radiation efficiency for the Brune spectrum (Eq. (14.32)) using the following steps. The Fourier transform of the displacement field of a seismogram (Eq. (14.32)) can be written $\dot{M}_0(f) = \frac{M_0}{1+(f/f_0)^2}$. Given Parseval's relation $\int_{-\infty}^{\infty} |u(t)|^2 \mathrm{d}t = \int_{-\infty}^{\infty} |u(f)|^2 \mathrm{d}f$ and the radiated energy (Eq. (15.47)),

(a) perform the energy integral in the frequency domain and show that

$$\frac{E_R^S}{M_0} = \frac{\pi^2 M_0 {f_0}^3}{5\rho V_S^5};$$

15.6 Equation (15.49) treats the sliding frictional block with displacement $y = y_e(1 - \cos(\omega t))$ and a moment acceleration that goes as $\ddot{M}_0 = 2\mu S\ddot{y} = 2\mu S\omega^2 y_e \cos\omega t$. That is, the block acceleration is

$$\ddot{y} = \omega^2 y_e \cos\omega t = \frac{4\pi^2}{T^2} y_e \cos\omega t, \text{for } 0 < t < T/2.$$

Note that the duration of sliding is *T*/2, where *T* is the period of the cosine. Now consider an alternative step-model in which

$$\ddot{y} = a_0, \;\; 0 < t \leq T/4,$$

$$\ddot{y} = -a_0, \;\; T/4 < t \leq T/2,$$

$$y_{\max} = 2y_e,$$

where values are adjusted so that the same moment is developed during the same time interval, but the acceleration is a positive step followed by negative step function that brings the block to rest over the same time interval.

(a) Recalling that the distance travelled, $y = 1/2\ddot{y}t^2$, find the acceleration a_0 by ensuring that the block travels $2y_e$.

(b) Compare the maximum accelerations in the two cases (cosine and step functions).

(c) Now integrate $\int_0^{\frac{T}{2}} \ddot{y}^2 \mathrm{d}t$ in the two cases (noting that $\int \cos^2 x \mathrm{d}x = \frac{x}{2} + \frac{1}{4}\sin(2x)$).

(d) The seismic and radiation efficiencies for the frictional blocks are 0.5% and 2.5%, respectively. By obtaining the ratios of the integrals in (c), calculate the efficiencies for the step-function case analysed in this problem.

(e) Are the efficiencies larger or smaller than for the sliding block model? Why? Is this model realistic?

16.1 Show that Eq. (16.22) can be derived by considering wave energies instead of amplitudes. What is the meaning of the possible alternative signs of R?

16.2 Consider a refracted ray that crosses a plane interface separating layers of velocities V_1 and V_2. Show by considering travel times of adjacent rays that Snell's law (Eq. (16.6)) is equivalent to Fermat's principle that the travel time for a seismic wave between source and receiver along a ray path is stationary (normally a minimum) with respect to adjacent paths, that is, the wave takes the fastest path.

16.3 Derive Eq. (16.75) by substitution of Eq. (16.71) for u in Eq. (16.74) and solving for the coefficients a_n in that equation by taking a Fourier transform and using Eq. (16.72) to resolve individual coefficients.

16.4 Given Eq. (16.35) for an S-wave incident on a boundary:

$$u^{\mathrm{inc}} = [u_x, u_z] = [\cos j, \sin j] \exp[\mathrm{i}\omega(px - \eta z - t)],$$

show that

(a) the reflected P-wave is

$$\begin{aligned} u^{\mathrm{refl}} &= [u_x, u_z] \\ &= R_{\mathrm{SP}}[\sin i, \cos i] \exp[\mathrm{i}\omega(px + \xi z - t)] \end{aligned}$$

(b) the reflected S-wave is

$$\begin{aligned} u^{\mathrm{refl}} &= [u_x, u_z] \\ &= R_{\mathrm{SS}}[\cos j, -\sin j] \exp[\mathrm{i}\omega(px + \eta z - t] \end{aligned}$$

(c) and Eq. (16.38), i.e.,

$$-2pV_{\mathrm{P}}V_{\mathrm{S}}\xi R_{\mathrm{SP}} + (1 - 2V_{\mathrm{S}}^2 p^2)(1 - R_{\mathrm{SS}}) = 0,$$

$$-(1 - 2V_{\mathrm{S}}^2 p^2)R_{\mathrm{SP}} + 2V_{\mathrm{S}}^3 p\eta/V_{\mathrm{P}}(1 + R_{\mathrm{SS}}) = 0,$$

follows from setting tractions on the surface equal to zero.

16.5 The Slichter mode of free oscillation is an oscillation, or rotation, of the inner core about the centre of the Earth. In principle, it provides the most precise observation of the density contrast between the inner and outer cores. Observations (by sensitive gravity meters) have been elusive, but it may now have been seen (Pagiatakis *et al.*, 2007). Consider a simple model of a uniform inner core, density ρ_{i}, in a non-rotating earth (with no splitting of the mode frequency), in a uniform outer core of density ρ_{o}. Calculate the restoring force for an off-centre displacement and allow for the fact that outer core material must move to accommodate inner core motion by assuming that the inertial mass of the motion is that of the inner core plus an equal volume of outer core material. Show that the period of the motion is

$$T = \sqrt{((3\pi/G\rho_{\mathrm{o}})(\rho_{\mathrm{i}} + \rho_{\mathrm{o}})/(\rho_{\mathrm{i}} - \rho_{\mathrm{o}}))}.$$

Assuming an inner–outer core density contrast of $820\,\mathrm{kg\,m^{-3}}$ (Masters and Gubbins, 2003), estimate the period (in hours). What are the significant approximations?

17.1 Consider a planet within which gravity is independent of depth (or radius). How does density vary with radius? Express the result in terms of mean density and total radius. (Note: gravity is approximately independent of depth in the Earth's mantle – Fig. 17.11(b). See also Problems 1.1(c) and 21.1.)

17.2 Deduce the velocities and thicknesses of plane horizontal layers that would give the following (hypothetical) first P-wave arrivals:

S (km)	T (s)	S (km)	T (s)
1	0.33	80	15.45
3	1.00	100	18.53
5	1.53	120	21.61
10	2.53	140	24.62
20	4.53	160	27.05
30	6.53	200	31.93
50	10.53	250	38.03
60	12.38	300	44.13

17.3 (a) A seismic refraction method is to be used to measure the thickness of sediment overlying igneous rock. If the P-wave velocities are estimated to be 2.5 km s^{-1} and 4.5 km s^{-1} and the postulated depth of sediment is 0.2 km, over what distance range should seismometers be placed to distinguish clearly between refractions from the igneous rock and direct P-waves from a surface explosion?

(b) A seismometer is buried at depth z in a uniform layer of P-wave velocity V_1 and thickness $z_1 > z$. There is another seismometer on the surface directly above it. Below z_1 is a thick layer of velocity $V_2 > V_1$. Both seismometers observe signals from a surface source that is sufficiently far away for the first arriving waves to have penetrated the deeper layer. What is the difference between first arrival times?

17.4 Derive Eq. (17.8). What is the maximum value of θ for which head waves could, in principle, be observed down-slope?

17.5 Consider a highly simplified model of the Earth, with three uniform layers: a core of radius 3500 km with P-wave speed $V = 9$ km s^{-1}, a lower mantle of thickness 2300 km (outer radius 5800 km) and $V = 12$ km s^{-1}, an upper mantle 600 km thick with $V = 9$ km s^{-1} (total radius 6400 km). What are the epicentral ranges of triplication of arrivals due to the mantle transition and the core shadow zone?

17.6 Show that in a flat Earth model in which seismic velocity increases linearly with depth, the seismic rays are circular arcs.

17.7 The P-wave speed at the base of the mantle (at 0.55 Earth radius) is 13.7 km s^{-1} and the velocity at the base of the crust (at negligible depth) is 6.4 km s^{-1}. What is the semi-angle of the cone in the crust to which seismic rays from the core are confined?

18.1 (a) Obtain an expression for the pressure at arbitrary radius r in a planet of uniform density, in terms of its mass, M, and radius, R.

(b) Show that the central pressure can be expressed in terms of the surface gravity, independently of radius or mass.

18.2 To draw attention to the increase in central pressure caused by an inward concentration of mass, consider a simple model, in which density varies with radius, r, as

$$\rho = a - br/R,$$

where $R = 6.371 \times 10^6$ m is the outer radius, $a = 13\,000$ kg m^{-3} is the central density and $b = 10\,000$ kg m^{-3}, the surface density, being $(a - b)$.

(a) Calculate the total mass of the model, to verify that it is very close to the mass of the Earth.

(b) Calculate the internal pressure as a function of radius and compare the central pressure with that of the Earth and of a uniform sphere of the same mass and radius.

(c) Calculate the moment of inertia coefficient, I/MR^2, for the model and compare it with a uniform sphere and the Earth. (See also the models in Problems 1.1 and 1.2.)

18.3 (a) Show that Eq. (18.21) with $m = 2$, $n = 4$ and all higher terms zero gives the second-order Birch finite strain equation,

$$P = \frac{3}{2}K_0\left(\frac{\rho}{\rho_0}\right)^{5/3}\left[\left(\frac{\rho}{\rho_0}\right)^{2/3} - 1\right],$$

where K_0 is the bulk modulus at $P = 0$.

(b) Show that, for this equation, $K'_0 = (dP/dK)_{P=0} = 4$.

18.4 A harmonic solid would be one in which the potential energy of an atomic displacement from equilibrium separation from a neighbour, r_0, is proportional to the square of displacement,

$$\phi(r) = A(r - r_0)^2,$$

so that atomic oscillations are sinusoidal or harmonic. Using this, with Eqs. (18.15) to (18.20), obtain expressions for

(a) P and K in terms of K_0 and (ρ/ρ_0),

(b) $K' = dK/dP$, in terms of (ρ/ρ_0), showing that $K'_0 = 1$,

(c) the Grüneisen parameter according to free volume-type theories (Eq. (19.39)) in terms of f and (ρ/ρ_0). (Note that γ and therefore also the thermal expansion coefficient are negative for a harmonic solid.)

19.1 For an ideal gas the familiar equation of state (n moles) is $PV = nRT$ and the principal specific heats are related by $C_P - C_V = R/m$ for atomic mass m.

(a) Show that for an ideal gas the Grüneisen parameter is $\gamma = C_P/C_V - 1$.
(Note: we are not using the ideal gas notation, in which γ represents C_P/C_V.)

(b) Use this result to show that the temperature variation under adiabatic compression is

$$T_1/T_2 = (V_2/V_1)^{\gamma}.$$

(c) Use these results to show that adiabatic compression is represented by

$$PV^{\gamma+1} = \text{constant}.$$

19.2 (a) Ignoring the differences between K_S and K_T and their pressure derivatives, use Eqs. (19.33) and (19.39), with elasticity data in Tables F.2 and F.3, to obtain rough estimates of the Grüneisen parameter at the top and bottom of the lower mantle and outer core.

(b) With C_V for the mantle by Eq. (19.3) and a multiplying factor 1.5 for the core, to allow for the electronic heat capacity, use the results in (a) with Eq. (19.1) to estimate the ranges of the expansion coefficient, α.

(c) With the results of (a) and (b) and the temperature ranges in Tables G.1 and G.2, estimate the ratio $K_S/K_T = (1 + \gamma\alpha T)$. This is a measure of the error in ignoring differences between K_S and K_T and their pressure derivatives (Eqs. (E.1) to (E.3) in Appendix E).

19.3 (a) Use Eq. (19.55) or (19.56) to derive the second term in Eq. (17.32), that is Birch's modification of the Williamson–Adams equation for a non-adiabatic gradient.

(b) Show that the factor $(1 - g^{-1} d\phi/dr)$, used by K. E. Bullen as a test for homogeneity in the Earth, is equivalent to dK/dP for a homogeneous, adiabatic layer. (Note that $\phi = K/\rho$ and r is radius in this problem.)

19.4 (a) Consider a model for a series of planets of different radii, R, with similar heat concentrations per unit volume, $\dot{q}$, and the same surface temperature, T_0. Assuming diffusive equilibrium, obtain expressions for the variation of central temperature with radius for each of two assumptions about conductivity:

(i) constant conductivity, κ,

(ii) $\kappa = AT^3$, where A is a constant,
(Case (ii) refers to radiative conductivity that can become important at high temperatures in minerals that are reasonably transparent. See Eq. (19.60).

(b) Consider constant κ, but $\dot{q}$ varying with radial distance r from the centre as $\dot{q} = \dot{q}_0(r/R)^l$, where l is a constant, independent of R, and $\dot{q}_0$ is the surface value of $\dot{q}$, which varies with R in such a way that the average heat generation per unit volume is independent of R. Find the central temperature as a function of R.

19.5 In spherical symmetry it is convenient to write the thermal diffusion equation with no internal heat sources (Eq. (20.1)) in terms of radius, r,

$$\partial T/\partial t = (\eta/r^2)\partial/\partial r(r^2 \partial T/\partial r).$$

Consider a sphere of radius R and constant diffusivity, η, with an initial temperature profile $T_0(r)$ which decays exponentially,

$$T(r) = T_0(r)\exp(-t/\tau).$$

Calculate the thermal relaxation time, τ, in terms of R and η, and the functional form of $T(r)$ that does not change with time.

19.6 Using the Lindemann relationship, $T_M \propto V^{2/3}{\theta_D}^2$, with the Debye approximation for γ (Eq. (19.29)), derive Eq. (19.43).

19.7 What lithospheric temperature gradient gives compensation of temperature and pressure effects on density, so that $d\rho/dz = 0$? Assume uniform composition and apply Eq. (17.32).

19.8 Derive Eq. (19.52) from Eqs. (19.50) and (19.51).

19.9 Using Table F.2 for numerical integration and assuming $T = T_M = 5000$ K at the inner core boundary ($r = 1221.5$ km), estimate T and T_M at the core–mantle boundary ($r = 3480$ km), using Eqs. (19.19) and (19.52) respectively. The difference is a measure of the cooling required for complete core solidification. (Note: use Eq. (19.39) for γ with $f = 1.44$.)

20.1 The oceanic lithosphere shrinks with age, due to cooling, resulting in a progressive increase in ocean depth with age of the ocean floor (Fig. 20.2). Isostatic balance (Section 9.3) is maintained and the added load of sea water causes a subsidence of the lithosphere, additional to its shrinkage. If sea water density is $\rho_w = 1025$ kg m^{-3} and the lithospheric density is $\rho_m = 3350$ kg m^{-3}, show that the depth increase exceeds the shrinkage by the factor $(1 - \rho_w/\rho_m)$ $^1 = 1.437$, as used in Eq. (20.11).

20.2 (a) Assuming a thermal oscillation at the surface of the Earth, $T = T_0 \sin\omega t$, and that at depth z the temperature variation is given by Eq. (20.23), show by differentiation and substitution in Eq. (20.1) that α and β satisfy Eq. (20.24).

(b) If the annual surface temperature oscillation has a peak-to-peak amplitude of 30 K and the diffusivity of surface rocks is 1.3×10^{-6} m^2 s^{-1}, at what depth in the crust must an instrument be buried to be subjected to a peak-to-peak annual temperature oscillation of 10^{-3} K?

20.3 Show that, at depth z, the temperature wave in Problem 20.2 has an oscillatory gradient of amplitude

$$\left(\frac{dT}{dz}\right)_{\max} = \sqrt{\frac{\omega}{\eta}}T_0 \exp\left(-\sqrt{\frac{\omega}{2\eta}}z\right).$$

20.4 Show that the temperature difference over the lower half of a borehole of depth z due to penetration of the thermal wave considered in Problems 20.2 and 20.3 oscillates with amplitude fT_0, where

$$f = \left[e^{-x}\left(e^{-x} + 1 - 2e^{-x/2}\cos\frac{x}{2}\right)\right]^{1/2},$$

$$x = \frac{z}{z^*} = \sqrt{\frac{\omega}{2\eta}}z,$$

z^* being the effective penetration depth (skin depth) of the wave. Show further that f has a maximum value of 0.347 at $x = 1.335$. If $z = 1000$ m and $\eta = 1.26 \times 10^{-6}$ m^2 s^{-1}, what is the corresponding period of the thermal wave?

20.5 Consider a region in which the crust has a total thickness z_0, with radioactive heat sources decreasing linearly with depth to zero at depth z_0. The heat flux from the mantle into the crust is a quarter of the surface heat flux and the temperature at the mantle–crust boundary is T'. Heat through the crust is by conduction only and the temperature profile is in equilibrium. Assuming that surface temperature is zero, obtain an expression for temperature as a function of depth z, down to z_0, in terms of T', z and z_0. By what factor does the gradient near the surface exceed the average crustal temperature gradient?

21.1 As in Problems 1.1(c) and 17.1, consider a self-gravitating body with density inversely proportional to radius. Compare the gravitational energy released by its formation with Eq. (21.2). What value of f is required?

21.2 If the heat flux from the Moon is in equilibrium with internal radioactivity, which has the same average concentration as the (mantle plus crust) of the Earth, what is the heat flux per unit area? (Relevant numerical details are given in Table 21.3.)

21.3 What is the total classical heat capacity of the Earth if the mantle (4×10^{24} kg) has a mean atomic weight of 21.1 and the core (2×10^{24} kg) has a mean atomic weight of 44.8? What average cooling would be required to produce the present heat flux for 4.5×10^9 years? Is this plausible? If so, why did Kelvin's age-of-the-Earth problem arise? (See Section 4.2.)

22.1 Consider a series of planets, as in Problem 19.4(a), except that they are convecting and maintain adiabatic gradients below their thin lithospheres, which have a common basal temperature, T_S. Assume that material compression follows Murnaghan's equation (see Eq. (18.37)) and that the Grüneisen parameter, γ, is a constant, so that Eq. (19.22) applies. Central pressure may be adopted from Problem 18.1(a) (which ignores the increase in density with depth or planet size). How does central temperature vary with radius?

22.2 Suppose that convected heat is derived uniformly from the whole volume of the mantle and the thermodynamic efficiency for each element of heat is proportional to the depth from which it is derived. The inner and outer radii are R_C and R_M and for heat from R_C the efficiency is $\eta_{\max}$. What is the average efficiency of whole mantle convection in terms of $\eta_{\max}$, R_C and R_M?

22.3 If the Grüneisen parameter of the mantle is assumed to be constant, $\gamma = 1.2$, and the relationship between incompressibility K and pressure P is

$$K = 2 \times 10^{11}\,\text{Pa} + 4P,$$

what is the thermodynamic efficiency of the convective transport of heat to the surface from the base of the mantle, where $P = 1.3 \times 10^{11}$ Pa?

22.4 Using the model of the core of Mercury in Section 24.8, with an inner core that has grown to 0.8 of the core radius, and assuming that a compositional density contrast of 5% was maintained between the inner and outer cores during inner core growth, estimate the total gravitational energy released during inner core growth. How significant would this be to the Mercury dynamo?

23.1 Is the decrease in moment of inertia of the Earth due to thermal contraction sufficient to cause an observable effect on rotation?

23.2 If the convected heat flux, $\dot{Q}$, from the mantle is proportional to the square root of the speed of convection, but the thermodynamic efficiency with which mechanical power is generated is independent of the speed, show that the effective viscosity is proportional to $\dot{Q}^{-3}$.

23.3 Use the heat balance equation (Eq. (23.14)) to set up a simplified numerical calculation for mantle cooling, by which radiogenic heat production in the mantle can be estimated. The targets of the calculation are the present value, $\dot{Q}_{R0}$, and the present cooling rate. Assume:

(i) radiogenic heat decays as a single exponential with a half life of 2×10^9 years,

(ii) convected heat, $\dot{Q}(T)$ varies with temperature by a simplified form of Eq. (23.21):

$$\dot{Q}(T) = \dot{Q}_0 \exp\left[\frac{2gT_M}{3(n+1)}\left(\frac{1}{T_0} - \frac{1}{T}\right)\right],$$

with

$T_0 = 1700$ K, the present value of T,
$T_M = 2500$ K, the value of T 4.5×10^9 years ago,
$\dot{Q}_0 = 32 \times 10^{12}$ W, the present convected heat loss,
$g = 25$, $n = 1$, $\phi_m = 7.4 \times 10^{27}\,\text{J K}^{-1}$.

The calculation is iterative. Start from the present conditions and by repeated trials integrate backwards in time to find the value of $\dot{Q}_{R0}$ that leads back to the required starting condition, $T = T_M$ at $t = -4.5 \times 10^9$ years. The value of $\dot{Q}_{R0}$ that you obtain will be less than $\dot{Q}_0$ and the difference gives the present cooling rate.

23.4 Taking the melting point of core alloy at the inner/outer core interface as 5000 K, and the temperature at the 660 km phase transition, 1950 K, from mineral physics, with elasticities from Appendix F and the Grüneisen parameter from Appendix G,

(a) extrapolate the temperature to the core–mantle boundary from both directions,

(b) if the difference between the two estimates is identified as the temperature increment across a thermal boundary layer 200 km thick, what is the average temperature gradient in it? If the thermal conductivity of the layer is 5 $\mathrm{W\,m^{-1}K^{-1}}$, what is the conducted heat? Is this consistent with heat conduction at the top of the core?

23.5 The mantle must have cooled much more than the core to leave a 1000 K thermal boundary layer at its base, and continuing faster mantle cooling is a feature of the thermal history, as presented in Chapter 23. Assuming this to be correct, calculate the maximum core cooling rate permitted by each of the three mantle cooling models, with different radiogenic heat, summarized by Figs. 23.2 and 23.3. What are the corresponding values of core-to-mantle heat flux if core radioactivity gives (a) 2tW, (b) zero? Note: take the mantle and core heat capacities to be given by Eqs. (21.7) and (21.20). The core value refers to cooling at the core–mantle boundary, but the mantle value refers to changes in the potential temperature and requires adjustment to an adiabatic extrapolation to the core–mantle boundary by the factor $\mathrm{T_P/\ T_{CMB}} = 1/1.64$.

24.1 Consider a spherical conductor with a pattern of currents that gives a uniform axial magnetic field throughout its volume. The field outside the sphere is equivalent to that of a dipole. Show that the field energy outside the sphere is half of the field energy within the sphere.

24.2 Show that, if the core is a sphere of radius R and uniform electrical resistivity, ρ_e, carrying a current that is a simple circulation about the axis, then

(a) if the current density is uniform it gives a dipole moment by Eq. (24.49) and therefore a power dissipation by Eq. (24.50),

(b) if the current density is proportional to radial distance, r, from the axis,

$$i = i_0 r/R,$$

then the magnetic moment is

$$m = (4\pi/15) i_0 R^4$$

and the ohmic dissipation is

$$-\frac{\mathrm{d}E}{\mathrm{d}t} = \frac{15}{2\pi}\frac{m^2 \rho_e}{R^5}.$$

This is the minimum possible dissipation for specified m.

24.3 Assuming the field strength in the core to be uniform with strength $2B_0$, where B_0 is the field on the equator just above the core surface, and to be related to dipole moment by Eq. (24.9), estimate the total field energy in terms of dipole moment (using the result of Problem (24.1)). By combining the resulting expression with the dissipation obtained in Problem 24.2(a), estimate the free decay time of the core currents and compare with the result from Eq. (24.43). (Note that this is only a rough calculation because the field is not uniform for the assumed current patterns.) Assume a uniform core resistivity of 4μΩ m.

24.4 What is the electromagnetic skin depth of the core for a magnetic fluctuation with (a) a 1000-year period? (b) a 10-year period? (Static layers at the top of the core have been considered. Their thickness is limited by the fact that secular variation gets through. These periods are extremes for the secular variation of harmonic terms in Fig. 24.4.)

24.5 Consider a model of the core rotating coherently about an axis at a small angle ϕ to the mantle axis, as in Problem 7.3. What is the value of ϕ that would explain the westward drift?

24.6 The gravitational energy released by adding to the inner core a mass dm of material that is more dense than the outer core, from which it is taken, and allowing the excess of light material to mix into the outer core, is given by Eq. (22.35). Obtain the corresponding expression for the energy released by mixing into the outer core a similar excess mass of material that is more dense than the outer core and is deposited from the mantle. By what factor does this energy differ from that given by Eq. (22.35)?

24.7 For a geocentric dipole field of strength given by Eq. (24.10), show that

(a) the rms surface field strength is $\sqrt{2}\mathrm{B}_0$,

(b) the mean (unsigned) field strength is $1.38\mathrm{B}_0$.

25.1 If the fraction of intrusive igneous rocks that have self-reversing remanent magnetism is f_1 and the fraction of the intruded rocks that self-reverse is f_2, show that the fraction of the intrusive rocks that disagree in polarity with their baked contacts is

$$F = (f_1 + f_2 - 2f_1f_2).$$

Observations suggest $F \approx 0.02$. If $f_1 = f_2 = f$, what are the possible values of f?

25.2 As mentioned in Section 25.5, paleointensity measurements use natural thermoremanence of igneous rocks, with the assumption that blocking temperatures are independent of cooling rate. This cannot be quite true and natural cooling rates may be $\sim 10^6$ times slower than laboratory cooling. To estimate the magnitude of the resulting error, calculate first the difference in E/kT_B for the two cooling rates by Eq. (25.2) and then use Fig. 25.2 to estimate the variation in spontaneous magnetization of magnetite with temperature at a blocking temperature $T_\mathrm{B} = 0.9\theta_\mathrm{C}$. Given that $E \propto m_\mathrm{s}^2$, calculate the variations in m_s and T_B for the different cooling rates. The values so calculated correspond to μ and T in Eq. (25.4). Assuming $\mu B/kT \ll 1$, this equation gives the magnetization at the blocking temperature, which is stronger at the lower blocking temperature. But cooling to low temperature causes a further increase in m_s to m_0 and this increase is less for the slower cooling because of the higher initial value of m_s. Therefore, only the T_B factor remains to cause an error in paleointensity estimates. How big is this error for a 10^6:1 difference in cooling rates?

25.3 If polar wander occurs at about 0.3° per million years, and the rms scatter of individual measurements of apparent instantaneous pole positions, caused by geomagnetic secular variation, is 15°, how many suitably averaged measurements are required to establish with 95% confidence that a pole position differs from one 30 million years earlier?

25.4 Viscous remanent magnetization (VRM) is the slow development of magnetic remanence at a fixed temperature. The process is not fundamentally different from thermoremanence, but involves blocking temperatures that are at ambient temperature on the relevant time scale. VRM induced by a field (or its decay after removal from a field) has often been reported to vary as the logarithm of time. This is not characteristic of grains all with the same activation energy, E, which would show an exponential relaxation towards equilibrium. Noting the sharpness of the blocking condition for individual grains, that is a small variation in T has a big effect on relaxation time, τ, use Eq. (25.1) or (25.2) to show that an assembly of grains with a uniform distribution of values of E can give a $\log t$ variation in VRM. (For a detailed discussion of VRM see Dunlop and Özdemir, 1997, Chapter 10.)

25.5 For reasons discussed in Section 25.4, we now discount the possibility that, immediately after a geomagnetic reversal, further reversals are inhibited. We may, not unreasonably, suggest that the opposite is the case, and that further reversals are more likely. For what may appear to be a rather subtle reason, this could be mistaken for inhibition in the reversal record. Why?

26.1 Variations in insolation (energy received from the Sun) are caused by regular variations in orbital characteristics (the Milankovitch cycles, discussed in Section 26.4). Assuming orbital angular momentum to be conserved, calculate the variation in annual average insolation due to the variation in orbital eccentricity *e*. (Equations in Appendix B are helpful to solution of this problem. A simple method is to calculate the total energy received on one complete orbit and then relate the orbital period to *e*.)

26.2 What is the effect on the Earth's orbit of friction of the solar tide? Could this have a significant influence on the global climate?

26.3 Calculate the effect on the length of the day of a 2 mm per year rise in sea level, assuming that it is due to (a) melting of the polar caps, and (b) thermal expansion of the surface layers of the oceans.

26.4 What is the required net heat input for each of cases (a) and (b) in Problem 26.3?

26.5 If the total energy use in Table 26.1 were derived entirely from tidal power stations, by what factor would the rate of change in the length of day increase?

References

Abercrombie, R. E. and Brune, J. N., 1994, Evidence for a constant *b*-value above magnitude 0 in the southern San Andreas, San Jacinto and San Miguel fault zones, and the Long Valley caldera, California. *Geophys. Res. Lett.* **21**: 1647–1650.

Abercrombie, R. and Leary, P., 1993, Source parameters of small earthquakes recorded at 2.5 km depth, Cajon Pass, southern California: implications for earthquake scaling. *Geophys. Res. Lett.* **20**: 1511–1514.

Abercrombie, R. E. and Rice, J. R., 2005, Can observations of earthquake scaling constrain slip weakening? *Geophys. J. Int.* **162**: 406–426.

Acton, G., Yin, Q.-Z., Verosub, K. L., Jovane, L., Roth, A., Jacobsen, B. and Ebel, D. S., 2007, Micromagnetic coercivity distributions and interactions in chondrules with implications for paleointensities of the early Solar System. *J. Geophys. Res.* **112**: B03S90. doi: 10.1029/2006JB004655.

Aharonson, O., Zuber, M. T. and Solomon, S. C., 2004, Crustal remanence in an internally magnetized non uniform shell: a possible source for Mercury's magnetic field. *Earth Planet. Sci. Lett.* **218**: 261–268.

Ahrens, T. J., ed., 1995a, *A Handbook of Physical Constants, 1: Global Earth Physics*. Washington: AGU.

Ahrens, T. J., ed., 1995b, *A Handbook of Physical Constants, 2: Mineral Physics and Crystallography*. Washington: AGU.

Ahrens, T. J., ed., 1995c, *A Handbook of Physical Constants, 3: Rock Physics and Phase Relations*. Washington: AGU.

Aki, K., 1969, Analysis of the seismic coda of local earthquakes as scattered waves. *J. Geophys. Res.* **74**: 615–631.

Aki, K. and Richards, P. G., 2002, *Quantitative Seismology*, second edn. Sausalito, CA: Science Books.

Alfè, D., Gillan, M. J. and Price, G. D., 2002, Composition and temperature of the Earth's core constrained by combining *ab initio* calculations and seismic data. *Earth Plan. Sci. Lett.* **195**: 91–98.

Alfvén, H., 1954, *The Origin of the Solar System*. Oxford: Clarendon Press.

Allègre, C. J., Poirier, J.-P. Humber, E. and Hofmann, A. W., 1995, The chemical composition of the Earth. *Earth Plan. Sci. Lett.* **134**: 515–526.

Allen, C. W., 1973, *Astrophysical quantities*, third edn. London: Athlone Press.

Alterman, Z., Jarosch, H., and Pekeris, C. L., 1959, Oscillations of the Earth. *Proc. Roy. Soc. Lond.* **A 252**: 80–95.

Alvarez, L. W., Alvarez, W., Asaro, F. and Michel, F. V., 1980, Extraterrestrial cause of the Cretaceous–Tertiary extinction. *Science* **208**: 1095–1108.

Anders, E., 1964, Origin, age and composition of meteorites. *Space Sci. Rev.* **3**: 583–714.

Anderson, E. M., 1905, Dynamics of faulting. *Trans. Edinburgh Geol. Soc.* **8**: 387–402.

Anderson, E. M., 1936, The dynamics of the formation of cone-sheets, ring-dykes, and caldron-subsidences, *Proc. Roy. Soc. Edin.* **56**: 128–156.

Anderson, J. D., Laing, P. A., Lau, E. L., Liu, A. S., Nieto, M. M. and Turyshev, S. G., 1998, Indication, from Pioneer 10/11, Galileo, and Ulysses data of an apparent anomalous, weak, long-range acceleration. *Phys. Rev. Lett.* **81**: 2858–2861.

Anderson, O. L., 1995, *Equations of State of Solids for Geophysics and Ceramic Science*. New York: Oxford University Press.

Anderson, O. L. and Isaak, D. G., 1995, Elastic constants of minerals at high temperature. In Ahrens (1995b), pp. 64–97.

Anderson, O. L. and Zou, K., 1990, Thermodynamic functions and properties of MgO at high compression and high temperature. *J. Phys. Chem. Ref. Data* **19**: 69–83.

Aoyama, Y. and Naito, I., 2001, Atmospheric excitation of the Chandler wobble, 1983–1998. *J. Geophys. Res.* **106**: 8941–8954.

Archer, C. L. and Jacobson, M. Z., 2005, Evaluation of global wind power. *J. Geophys. Res.* **110**: d12110, doi:10.1029/2004JD005462.

Atkinson, B. K., 1982, Subcritical crack propagation in rocks: theory, experimental results and applications. *J. Struct. Geol.* **4**: 41–56.

Balling, R. C. and Cerveny, R. S., 1995, Impact of lunar phase on the timing of global and latitudinal tropospheric temperature maxima. *Geophys. Res. Lett.* **22**(23): 3199–3201.

Bard, B., Hamelin, B., Fairbanks, R. G. and Zindler, A., 1990, Calibration of the ^{14}C timescale over the past 30,000 years using mass spectrometric U-Th ages from Barbados corals. *Nature* **345**: 405–410.

Barton, C. E., 1989, Geomagnetic secular variation: direction and intensity. In James (1989), pp. 560–577.

Bass, J. D., 1995, Elasticity of minerals, glasses and melts. In Ahrens (1995b), pp. 45–63.

Benz, H. M. and Vidale J. E., 1993, Sharpness of upper-mantle discontinuities determined from high-frequency reflections. *Nature* **365**: 147–150.

Berger, A., 1988, Milankovitch theory and climate. *Rev. Geophys.* **26**: 624–657.

Berger, A. and Loutre, M. F., 1992, Astronomical solutions for paleoclimate studies over the last 3 million years. *Earth Plan. Sci. Lett.* **111**: 369–382.

Bernatowicz, T. J. and Walker, R. M., 1997, Ancient stardust in the laboratory. *Physics Today* **December 1997**: 26–32.

Bi, Y., Tan. H. and Jin, F., 2002, Electrical conductivity of iron under shock compression up to 200GPa. *J. Phys. Condensed Matter* **14**; 10849–10854.

Bina, C. R. and Helffrich, G. R., 1994, Phase transition Clapeyron slopes and transition zone seismic discontinuity topography. *J. Geophys. Res.* **99**: 15853–15860.

Birch, F., 1952, Elasticity and constitution of the Earth's interior. *J. Geophys. Res.* **57**: 227–286.

Bird, P., 1978, Finite element modelling of lithosphere deformation: the Zagros collision orogeny. *Tectonophysics* **50**: 307–336.

Bird, P., 1998, Testing hypotheses on plate driving mechanisms with global lithosphere models including topography, thermal structure and faults. *J. Geophys. Res.* **103**(B5): 10115–10129.

Bird, P., 2003, An updated digital model of plate boundaries. *Geochemistry Geophysics Geosystems* **4**(3): 1027. doi: 10.1029/2002GLO16002.

Bird, P. and Kagan, Y. Y., 2004, Plate-tectonic analysis of shallow seismicity; apparent boundary width, beta, corner magnitude, coupled lithosphere thickness, and coupling in seven tectonic settings. *Bull. Seism. Soc. Am.* **94**: 2380–2399.

Blackett, P. M. S., 1952, A negative experiment relating to magnetism and the Earth's rotation. *Phil. Trans. Roy. Soc. Lond.* **A245**: 309–370.

Bloxham, J., 2002, Time-independent and time-dependent behaviour of high-latitude flux bundles at the core–mantle boundary. *Geophys. Res. Lett.* **29**(18), doi:10.1029/2001GLO14543.

Bloxham, J., Gubbins, D. and Jackson, A., 1989, Geomagnetic secular variation. *Phil. Trans. Roy. Soc. Lond.* **A329**: 415–502.

Blyth, A. E., Burbank, D. W., Farley, K. A. and Fielding, E. J., 2000, Structural and topographic evolution of the central Transverse Ranges, California, from apatite fission track, (U-Th)/He and digital elevation model analyses. *Basin Research* **12**: 97–114.

Boehler, R., 1993, Temperatures in the Earth's core from melting point measurements of iron at high static pressures. *Nature* **363**: 534–536.

Boehler, R., 2000, High pressure experiments and the phase diagram of lower mantle and core materials. *Rev. Geophys.* **38**: 221–245.

Boness, D. A., Brown, J. M. and McMahan, A. K., 1986, The electronic thermodynamics of iron under Earth's core conditions. *Phys. Earth Planet. Inter.* **42**: 227–240.

Bonner, J. L., Blackwell, D. D. and Herrin, E. T., 2003, Thermal constraints on earthquake depths in California. *Bull. Seism. Soc. Am.* **93**: 2333–2354.

Born, M. and Wolf, E., 1965, *Principles of Optics*. Oxford: Pergamon.

Boschi, L. and Dziewonski, A. M., 2000, Whole Earth tomography from delay times of P, PcP, PKP phases: lateral heterogeneities in the outer core, or radial anisotropy in the mantle? *J. Geophys. Res.* **105**: 25567–25594.

Bottke, W. F., Vokrouhlický, D., Rubincam, D. P. and Nesvorný, D., 2005, The Yarkovsky and YORP effects: implications for asteroid dynamics. *Ann. Rev. Earth Plan. Sci.* **34**: 157–191.

Bouhifd, M. A., Gautron, L., Bolfan-Casanova, N., Malavergne, V., Hammouda, T., Andrault, D. and Jephcoat, A. P., 2007, Potassium partitioning into molten iron alloys at high pressure: implications for Earth's core. *Phys. Earth Planet. Inter.* **160**: 22–33.

Bowman, D. D. and King, G. C. P., 2001, Accelerating seismicity and stress accumulation before large earthquakes, *Geophys. Res. Lett.* **28**: 4039–4042.

Boyet, M. and Carlson, R. W., 2005, Nd evidence for early (>4.53 Ga) global differentiation of the silicate earth. *Science* **309**: 5756–580.

Braginsky, S. I., 1991, Towards a realistic theory of the geodynamo. *Geophys. Astrophys. Fluid Dyn.* **60**: 89–134.

Braginsky, S. I., 1993, MAC-oscillations of the hidden ocean of the core. *J. Geomag. Geoelect.* **45**: 1517–1538.

Braginsky, S. I., 1999, Dynamics of the stably stratified ocean at the top of the core. *Phys. Earth Plant. Inter.* **111**: 21–34.

Braginsky, S. I. and Roberts, P. H., 1995, Equations governing convection in the Earth's core and the geodynamo. *Geophys. Astrophys. Fluid Dynam.* **79**: 1–97.

Brennan, B. J. and Smylie, D. E., 1981, Linear viscoelasticity and dispersion in seismic wave propagation. *Rev. Geophys. Space Phys.* **19**: 233–246.

Bridgman, P. W., 1914, A complete collection of thermodynamic formulas. *Phys. Rev.* **3**: 273–281.

Bridgman, P. W., 1957, Effects of pressure on binary alloys, V and VI. *Proc. Am. Acad. Arts Sci.* **84**: 131–216.

Brown, M. E., Trujillo, C. and Rabinowitz, D., 2004, Discovery of a candidate inner Oort cloud planetoid. *Astrophys. J.* **617**: 645–649.

Brown, M. E., Trujillo, C. and Rabinowitz, D., 2005, Discovery of a planetary-sized object in the scattered Kuiper belt. *Astrophys. J.* **635**: L97–L100.

Brune, J. N., 1968, Seismic moment, seismicity, and rate of slip along major fault zones. *J. Geophys. Res.* **83**: 777–784.

Brune, J. N., 1970, Tectonic stress and the spectra of seismic shear waves from earthquakes. *J. Geophys. Res.* **75**: 4997–5009.

Budner, D. and Cole-Dai, J., 2003, The number and magnitude of large explosive volcanic eruptions between 904 and 1865AD: quantitative evidence from a new South Pole ice core. In Robock, C. and Oppenheimer, C. (eds.), 2003, *Volcanism and the Earth's Atmosphere*. Washington: American Geophysical Union, pp. 165–176.

Buffett, B. A., 1992, Constraints on magnetic energy and mantle conductivity from the forced nutations of the Earth. *J. Geophy. Res.* **97**: 19581–19597.

Buffett, B. A., 1996, A mechanism for decade fluctuations in the length of day. *Geophys. Res. Lett.* **23**: 3803–3806.

Buffett, B. A., 1997, Geodynamic estimates of the viscosity of the Earth's inner core. *Nature* **388**: 571–573.

Bukowinsky, M. S. T. and Knopoff, L., 1977, Physics and chemistry of iron and potassium. In Manghnani, M. H. and Akimoto, S. (eds.), 1977, *High Pressure Research: Applications in Geophysics*. New York: Academic Press.

Bullard, E. C., Everett, J. E. and Smith, A. G., 1965, The fit of the continents around the Atlantic. *Phil. Trans. Roy. Soc. Lond.* **A258**: 41–51.

Bullard, E. C., Freedman, C., Gellman, H. and Nixon, J., 1950, The westward drift of the Earth's magnetic field. *Phil. Trans. Roy. Soc. Lond.* **A243**: 67–92.

Bullard, E. C. and Gellman, H, 1954, Homogeneous dynamos and geomagnetism. *Phil. Trans. Roy. Soc. Lond.* **A247**: 213–255.

Bullen, K. E., 1975, *The Earth's Density*. London: Chapman and Hall.

Bullen, K. E. and Bolt, B. A., 1985, *An Introduction to the Theory of Seismology*. Cambridge: Cambridge University Press.

Burchfield, J. D., 1975, *Lord Kelvin and the Age of the Earth*. New York: Science History Publications.

Busse, F. H., 2002, Convective flows in rapidly rotating spheres and their dynamo action. *Phys. Fluids* **14**: 1301–1314.

Byerlee, J. D., 1978, Friction in rocks. *Pure Appl. Geophys.* **116**: 615–626.

Cagniard, L., 1939, *Réflexion et réfraction des ondes séismique progressives*. Paris: Gauthier-Villard.

Cagniard, L., 1962, *Reflection and Refraction of Progressive Seismic Waves*. Translation by E. A. Flinn and C. H. Dix. New York: McGraw-Hill.

Cain, J. C., Wang, Z., Schmitz, D. R. and Meyer, J., 1989, The geomagnetic model spectrum for 1980 and core-crustal separation. *Geophys. J. Int.* **97**: 443–447.

Cande, S. and Kent, D. V., 1995, Revised calibration of the geomagnetic polarity time scale for the Late Cretaceous and Cenozoic. *J. Geophys. Res.* **100**: 6093–6095.

Canup, R. M. and Asphaug, E., 2001, Origin of the Moon in a giant impact near the end of the Earth's formation. *Nature* **412**: 708–712.

Carlson, R. W., Hilde, T. W. C. and Uyeda, S., 1983, The driving mechanism of plate tectonics: relation to the age of the lithosphere at trenches. *Geophys. Res. Lett.* **10**: 297–300.

Carter, W. E., 1989, Earth orientation. In James (1989), pp. 231–239.

Cazenave, A., 1995, Geoid, topography and distribution of landforms. In Ahrens (1995a), pp. 32–39.

Chambat, F. and Valette, B., 2005, Earth gravity up to second order in topography and density. *Phys. Earth Planet. Inter.* **151**: 89–106.

Chao, B. F., 1995, Anthropogenic impact on global geodynamics due to reservoir water impoundment. *Geophys. Res. Lett.* **22**: 3529–3532.

Chao, B. F., Au, A. Y., Boy, J.-P. and Cox, C. M., 2003, Time-variable gravity signal of an anomalous redistribution of water mass in the extratropic Pacific during 1998–2002. *Geochem. Geophys. Geosyst.* **4** (11):1096, doi:10.1029/2003GG000589.

Chao, B. F., Rodenburg, E., Sahagian, D. L., Jacobs, D. K. and Schwartz, F. W., 1994, Man made lakes and sea level rise. *Nature* **370**: 258.

Chapman, S. and Bartels, J., 1940, *Geomagnetism*. London: Oxford University Press.

Chow, T. J. and Patterson, C. C., 1962, The occurrence and significance of lead isotopes in pelagic sediments. *Geochim. Cosmochim. Acta* **26**: 263–308.

Christodoulidis, D. C., Smith, D. E., Williamson, R. G. and Klosko, S. M., 1988, Observed tidal braking in the Earth/Moon/Sun system. *J. Geophys. Res.* **93**: 6216–6236.

Clark, S. P., 1957, Radiative transfer in the Earth's mantle. *Trans. Am. Geophys. Un.* **38**: 931–938.

Clauser, C. and Huenges, E., 1995, Thermal conductivity of rocks and minerals. In Ahrens (1995c), pp. 105–126.

Clayton, R. N., 2002, Self-shielding in the solar nebula. *Nature* **415**; 860–861.

Coblentz, D. D., Zhou, S., Hillis, R. R., Richardson, R. M. and Sandiford, M., 1998, Topography, boundary forces, and the Indo-Australian intraplate stress field. *J. Geophys. Res.* **103**(B1): 919–931.

Cohen, B.A., Swindle, T.D. and Kring, D.A., 2000, Support for the lunar cataclysm hypothesis from lunar meteorite impact melt ages. *Science* **290**: 1754–1756.

Connerney, J.E.P. *et al.* (10 authors), 1999, Magnetic lineations in the ancient crust of Mars. *Science* **284**; 794–798.

Constable, C.G. and Parker, R.L., 1988, Statistics of the secular variation for the past 5 my. *J. Geophys. Res.* **93**: 11569–11581.

Courboulex, F., Singh, S.K., Pacheco, F. and Ammon, C.J., 1997, The 1995 Colima-Jalisco, Mexico, earthquake (Mw8): a study of the rupture process. *Geophy. Res. Lett.* **24**(9): 1019–1022.

Courtillot, V., 1999, *Evolutionary Catastrophes: the Science of Mass Extinction*. Cambridge: Cambridge University Press.

Courtillot, V. and LeMouël, J.L., 1984, Geomagnetic secular variation impulses. *Nature* **311**: 709–716.

Cox, A., 1973, *Plate Tectonics and Geomagnetic Reversals*. San Francisco: W.H. Freeman.

Cox, A., Doell, R.R. and Dalrymple, G.B., 1963, Geomagnetic polarity epochs and pleistocene geochronometry. *Nature* **198**: 1049–1051.

Cox, A. and Hart, R.B., 1986, *Plate Tectonics: How it Works*. Palo Alto: Blackwell Scientific Publications.

Cox, C.M. and Chao, B.F., 2002, Detection of a large scale mass redistribution in the terrestrial system since 1998. *Science* **297**: 831–833.

Crampin, S., 1977, A review of the effects of anisotropic layering on the propagation of seismic waves. *Geophys. J. R. Astr. Soc.* **49**: 9–27.

Creager, K.C., 1997, Inner core rotation from small scale heterogeneity and time-varying travel times. *Science* **278**: 1284–1288.

Creer, K.M. and Tucholka, P., 1982, Secular variation as recorded in lake sediments: a discussion of North American and European results. *Phil. Trans. Roy. Soc. Lond* **A306**: 87–102.

Curie, P., 1894, Sur la symétrie dans les phénomènes physiques, symétrie d'un champ electrique et d'un champ magnétique. *J. de Phys.* (Paris) **3**: 393–415.

Dahlen, F.A., Hung, S.-H. and Nolet, G., 2000, Fréchet kernels for finite-frequency traveltimes – I. Theory. *Geophys. J. Int.* **141**: 157–174.

Dahlen, F.A. and Tromp, J., 1998, *Theoretical Seismology*. Princeton: Princeton University Press.

Dainty, A.M., 1990, Studies of coda using array and three-component processing. *Pure Appl. Geoph.* **132**: 221–244.

Dainty, A., 1995, The influence of seismic scattering on monitoring. In Husebye, E.S. and Dainty, A. (eds.), *Monitoring a Comprehensive Test Ban Treaty*. Dordrecht: Kluwer, pp. 663–688.

Dalrymple, G.B. and Ryder, G., 1993, ^{40}Ar/^{39}Ar age spectra of Apollo 15 impact melt rocks by laser step-heating and their bearing on the history of lunar basin formation. *J. Geophys. Res.* **98**(E7): 13085–13096.

Dalrymple, G.B. and Ryder, G., 1996, Argon-40/argon-39 age spectra of Apollo 17 highlands breccia samples by laser step heating and the age of the Serenitatis basin. *J. Geophys. Res.* **101**(E11): 26069–26084.

Das, S., 1981, Three-dimensional spontaneous rupture propagation and implications for the earthquake source mechanism. *Geophys. J. Roy. Astr. Soc.* **67**: 375–393.

Davis, D., Suppe, J. and Dahlen, F.A., 1983, Mechanics of fold-and-thrust belts and accretionary wedges. *J. Geophys. Res.* **88**: 1153–1172.

Davis, P.M., 1983, Surface deformation associated with a dipping hydrofracture. *J. Geophys. Res.* **88**: 5826–5833.

Davis, P.M., 1986, Surface deformation due to inflation of an arbitrarily oriented triaxial ellipsoidal cavity in an elastic half-space with reference to Kilauea volcano, Hawaii. *J. Geophys. Res.* **91**: 7429–7430.

Davis, P.M., 2003, Azimuthal variation in seismic anisotropy of the Southern California uppermost mantle. *J. Geophys. Res.* **108**(B1): 2052. doi10.1029/2001JB000637,2003.

Davis, P.M., Rubenstein, J.L., Liu, K.H., Gao, S.S. and Knopoff, L., 2000, Northridge earthquake damage caused by geologic focusing of seismic waves. *Science* **289**: 1746–1750.

Dearden, E.W., 1995, Expansion formulae for first order partial derivatives of thermal variables. *Eur. J. Phys.* **16**: 76–79.

Degens, E.T. and Ross, D.A. (eds.), 1969, *Hot Brines and Recent Heavy Metal Deposits in the Red Sea*. New York: Springer.

Dehant, V., Creager, K.C., Karato, S.-I. and Zatman, S. (eds.), 2003, *Earth's Core: Dynamics, Structure, Rotation*. Geodynamics Series 31. Washington: American Geophysical Union.

DeHoop, A.T., 1960, Modification of Cagniard's method for solving seismic pulse problems. *Appl. Sci. Res.* **B8**: 349–356.

DeMets, C., Gordon, R.G., Argus, D.F. and Stein, S., 1990, Current plate motions. *Geophys. J. Int.* **101**: 425–478.

DeMets, C., Gordon, R.G., Argus, D.F. and Stein, S., 1994, Effect of recent revisions of the geomagnetic reversal time scale on estimates of current plate motions. *Geophys. Res. Lett.* **21**: 2191–2194.

DePaolo, D. J., 1981, Nd isotopic studies: some new perspectives on Earth structure and evolution. *EOS* (*Trans. Am. Geophys. Un.*) **62**: 137–140 (April 7, 1981).

Deuss, A. and Woodhouse, J., 2001, Seismic observations of splitting of the mid-transition zone discontinuity in Earth's mantle. *Science* **294**: 354–357.

Deuss, A., Woodhouse, J. H., Paulssen. H. and Trampert, J., 2000, The observation of inner core shear waves. *Geophys. J. Int.* **142**: 67–73.

Dieterich, J. H., 1979a, Modeling of rock friction 1, experimental results and constitutive equations. *J. Geophys. Res.* **84**: 2161–2168.

Dieterich, J. H., 1979b, Modeling of rock friction 2, simulation of preseismic slip. *J. Geophys. Res.* **84**: 2169–2175.

Dieterich, J., 1994, A constitutive law for rate of earthquake production and its application to earthquake clustering. *J. Geophys. Res.* **99**: 2601–2618.

Dobson, D. P., 2002, Self-diffusion in liquid Fe at high pressure. *Phys. Earth Planet. Inter.* **130**: 271–284.

Dobson, D. P. and Brodholt, J. P., 2000, The electrical conductivity and thermal profile of the Earth's mid mantle. *Geophys. Res. Lett.* **27**: 2325–2328.

Doornbos D. J. 1974, The anelasticity of the inner core. *Geophys. J. R. Astron. Soc.* **38**: 397–415.

Doornbos, D. J., 1992, Diffraction and seismic tomography. *Geophys. J. Int.* **108**: 256–266.

Doornbos, D. J. and Hilton, T., 1989, Models of the core–mantle boundary and the travel times of internally reflected core phases. *J. Geophys. Res.* **94**: 15741–15751.

Dragert, H. K., Wang, K. and James, T. S., 2001, A silent slip event on the deeper Cascadia subduction interface, *Science* **292**: 1525–1528.

Duffield, W. A., 1972, A naturally occurring model of global plate tectonics. *J. Geophys. Res.* **77**: 2543–2555.

Dugdale, J. S. and MacDonald, D. K. C., 1953, Thermal expansion of solids. *Phys. Rev.* **89**: 832–834.

Dunlop, D. J. and Özdemir, Ö., 1997, *Rock Magnetism: Fundamentals and Frontiers*. Cambridge: Cambridge University Press.

Dwight, H. B., 1961, *Tables of Integrals and Other Mathematical Data*, fourth edn. New York: MacMillan.

Dziewonski, A. M., 1984, Mapping the lower mantle: determination of lateral heterogeneity in P velocity up to degree and order 6. *J. Geophys. Res.* **89**: 5929–5952.

Dziewonski, A. M. and Anderson, D. L., 1981, Preliminary reference Earth model. *Phys. Earth Planet. Inter.* **25**: 297–356.

Dziewonski, A. M., Chou, T.-A. and Woodhouse, J. H., 1981, Determination of earthquake source parameters from waveform data for studies of global and regional seismicity. *J. Geophys. Res.* **86**: 2825–2852.

Earle, P. S. and Shearer, P. M., 2001, Observations of PKKP precursors used to estimate small scale topography on the core–mantle boundary. *Science* **277**: 667–670.

Eaton, J. P., Richter, D. H. and Ault, W. U., 1961, The tsunami of May 3, 1960, on the island of Hawaii. *Bull. Seism. Soc. Am.* **51**: 135–157.

Ekman, M., 1993, A concise history of the theories of tides, precession-nutation and polar motion (from antiquity to 1950). *Surveys in Geophys.* **14**: 585–617.

Eldridge, J. S., O'Kelly, G. D. and Northcutt, K. J., 1974, Primordial radioelement concentrations in rocks from the Taurus-Littrow. *Proc. Fifth Lunar Conference* (*Suppl. 5, Geochim. Cosmochim. Acta*) **2**: 1025–1031.

Elsasser, W. M., 1978, *Memoirs of a Physicist in the Atomic Age*. New York: Science History Publications and Bristol: Adam Hilger.

Eshelby, J. D., 1973, Dislocation theory for geophysical applications. *Phil. Trans. Roy. Soc. Lond* **A274**: 331–338.

Eymin, C. and Hulot, G., 2005, On core surface flows inferred from satellite magnetic data. *Phys. Earth Planet. Inter.* **152**: 200–220.

Falzone, A. J. and Stacey, F. D., 1980, Second order elasticity theory: explanation for the high Poisson's ratio of the inner core. *Phys. Earth Planet. Inter.* **21**: 371–377.

Farley, K. A., Vokrouhlický, D., Bottke, W. F. and Nesvorný, D., 2006, A late Miocene dust shower from the break-up of an asteroid in the main belt. *Nature* **439**: 295–297.

Fearn, D. R. and Loper, D. E., 1981, Compositional convection and stratification of the Earth's core. *Nature* **289**: 393–394.

Fegley, B., 1995, Properties and composition of the terrestrial oceans and of the atmospheres of the Earth and other planets. In Ahrens (1995a). pp. 320–345.

Felzer, K. R. and Brodsky, E. E., 2006, Decay of aftershock density with distance indicates triggering by dynamic stress. *Nature* **411**: 735–738.

Fisher, D. E., 1975, Trapped helium and argon and the formation of the atmosphere by degassing. *Nature* **256**: 113–114.

Fitzgerald, R., 2003, Isotope measurements firm up knowledge of Earth's formation. *Physics Today* **January 2003**: 16–18.

Flanagan, M. P. and Shearer, P. M, 1998, Global mapping of topography on transition zone velocity discontinuities by stacking *SS* precursors. *J. Geophys. Res.* **103**: 2673–2692.

Fleischer, R. L., Naeser, C. W., Price, P. B., Walker, R. M. and Maurette, M., 1965, Cosmic ray exposure ages of tektites by the fission track technique. *J. Geophys. Res.* **70**: 1491–1496.

Forsyth, D.W., 1975, The early structural evolution and anisotropy of the oceanic upper mantle. *Geophys. J. R. Astr. Soc.* **43**: 103–162.

Forte, A.M. and Mitrovica, J.X., 2001, Deep-mantle high-viscosity flow and thermochemical structure inferred from seismic and geodynamic data. *Nature* **410**: 1049–1056.

Fowler, W.A., 1961, Rutherford and nuclear cosmochronology. *Proc. Rutherford Jubilee Intern. Conf.*, ed. J.B. Birks, pp. 640–676. London: Heywood.

Furumura T. and Kennett B.L.N., 2005, Subduction zone guided waves and the heterogeneity structure of the subducted plate. *J. Geophys. Res.* **110**: B10302, doi:10.1029/2004JB003486.

Gessman, C.K. and Wood, B.J., 2002, Potassium in the Earth's core? *Earth Plan. Sci. Lett.* **200**: 63–78.

Gillet, P., Richet, P., Guyot, F. and Fiquet, G., 1991, High temperature thermodynamic properties of forsterite. *J. Geophys. Res.* **96**: 11805–11816.

Gilvarry, J.J., 1956, The Lindemann and Grüneisen laws. *Phys. Rev.* **102**: 308–316.

Glatzmaier, G.A. and Roberts, P.H., 1995a, A three-dimensional convective dynamo solution with rotating and finitely conducting inner core and mantle. *Phys. Earth Planet. Inter.* **91**: 63–75.

Glatzmaier, G.A. and Roberts, P.H., 1995b, A three-dimensional self consistent computer simulation of a geomagnetic field reversal. *Nature* **377**: 203–209.

Glatzmaier, G.A. and Roberts, P.H., 1996, Rotation and magnetism of the Earth's inner core. *Science* **274**: 1887–1891.

Goldstein, J.I. and Ogilvie, R.E., 1965, A re-evaluation of the iron-rich portion of the Fe-Ni system. *Trans. Metall. Soc. AIME* **233**: 2083–2087.

Goncharov, A.F., Struzhkin, V.V. and Jacobsen, S.D., 2006, Reduced radiative conductivity of low-spin (Mg, Fe)O in the lower mantle. *Science* **312**: 1205–1208.

Gordon, A.H., 1994, Weekdays warmer than weekends. *Nature* **367**: 325–326.

Gough, D.I. and Gough, W.I., 1970, Stress and deflection in the lithosphere near Lake Kariba. *Geophys. J. Roy. Astron. Soc.* **21**: 65–78.

Gradstein, F.M. *et al.* (40 authors), 2005, *A Geologic Time Scale 2004*. Cambridge: Cambridge University Press. (www.stratigraphy.org/gts.htm).

Grand, S.P. 2001, The implications for mantle flow from global seismic tomography. In *Integrated models of Earth structure and evolution, AGU Virtual Spring Meeting, 20 June 2001.* (www.agu.org/meetings/umeeting/.)

Grand, S.P., van der Hilst, R.D. and Widiyantoro, S., 1997, Global seismic tomography: a snapshot of convection in the Earth. *GSA Today* **7**: 1–7.

Gray, C.M., Papanastassiou, D.A. and Wasserburg, G.J., 1973, The identification of early condensates from the solar nebula. *Icarus* **20**: 213–239.

Gross, R.S., 2000, The excitation of the Chandler wobble. *Geophys. Res. Lett.* **27**: 2329–2332.

Gross, R.S., 2001, A combined length-of-day series spanning 1832–1997: LUNAR97. *Phys. Earth Planet. Inter.* **123**: 65–76.

Gubbins, D., 1977, Energetics of the Earth's core. *J. Geophys.* **43**: 453–464.

Gubbins, D., 1994, Geomagnetic polarity reversals: a connection with secular variation and core–mantle interaction? *Rev. Geophys.* **32**: 61–83.

Gubbins, D., 2003, Thermal core–mantle interactions: theory and observations. In Dehant *et al.* (2003), pp. 163–179.

Gung, Y.C. and Romanowicz, B., 2004, Q tomography of the upper mantle using three component long period waveforms. *Geophys. J. Int.* **157**: 813–830.

Gutenberg, B. and Richter, C.F., 1941, Seismicity of the Earth. *Geol. Soc. Am. Spec. Pap.* **34**: 1–131.

Haak, V. and Jones, A.G., 1997, Introduction to special section: the KTB deep drill hole. *J. Geophys. Res.* **102**(B8): 18175–18177.

Haddon, R.A.W., 1972, Corrugations on the CMB or transition layers between inner and outer cores? *Trans. Am. Geophys. Un.* **53**: 600.

Haddon, R.A.W. and Cleary, J.R., 1974, Evidence for scattering of seismic PKP waves near the mantle–core boundary, *Phys. Earth Planet. Int.* **8**: 211–234.

Haddon, R.A.W., Husebye, E.S. and King D.W., 1977, Origins of precursors to PP. *Phys. Earth Planet. Int.* **14**: 41–70.

Hager, B.H., 1984, Subducted slabs and the geoid: constraints on mantle rheology and flow. *J. Geophys. Res.* **89**: 6003–6015.

Hager, B.H. and Richards, M.A., 1989, Long wavelength variations in the Earth's geoid: physical models and dynamical implications. *Phil. Trans. Roy. Soc. Lond.* **A328**: 309–327.

Hale, C.J., 1987, The intensity of the geomagnetic field at 3.5 Ga: paleointensity results from the Komati Formation, Barberton Mountain Land, South Africa. *Earth Plan. Sci. Lett.* **86**: 354–364.

Halls, H.C., McArdle, N.J., Gratton. M.H. and Shaw, J., 2004, Microwave paleointensities from dyke chilled margins: a way to obtain long-term variations in geodynamo intensity for the last three billion years. *Phys. Earth Planet. Inter.* **147**: 183–195.

Han, D. and Wahr, J., 1995, The viscoelastic relaxation of a realistically stratified earth, and a further analysis of postglacial rebound. *Geophys. J. Int.* **120**: 287–311.

Harrison, T. M., Blichert-Toft, J., Müller, W., Albarede, F., Holden, P. and Mojzsis, S. J., 2005, Heterogeneous Hadean hafnium: evidence of continental crust at 4.4 to 4.5 Ga. *Science* **310**: 1947–1950.

Hart, R., Hogan, L. and Dymond, J., 1985, The closed system approximation for evolution of argon and helium in the mantle, crust and atmosphere. *Chem. Geol.* (*Isotope Geoscience Section*) **52**: 45–73.

Hartman, W. K., 2003, Megaregolith evolution and cratering cataclysm models – lunar cataclysm as a misconception (28 years later). *Meteoritics and Planetary Science* **38**: 579–593.

Hasegawa, A., 1989, Seismicity: subduction zone. In James (1989), pp. 1054–1061.

Hasegawa, A., Zhao, D., Shuichiro, H., Yamamoto, A. and Horiuchi, S., 1991, Deep structure of the northeastern Japan arc and its relationship to seismic and volcanic activity. *Nature* **352**: 683–689.

Hashin, Z. and Shtrikman, S., 1963, A variational approach to the elastic behaviour of multiphase materials. *J. Mech. Phys. Solids* **11**: 127–140.

Haskell, N. A., 1935, The motion of a fluid under a surface load, 1, *Physics* **6**: 265–269.

Haskell, N. A., 1969, Elastic displacements in the near-field of a propagating fault, *Bull. Seism. Soc. Am.* **59**: 865–908.

Hayatsu, A. and Wabaso, C. E., 1985, The solubility of rare gases in silicate melts and implications for K-Ar dating. *Chem. Geology* (*Isotope Geoscience Section*) **52**: 97–102.

Hearn, E. H., 2003, What can GPS data tell us about the dynamics of post-seismic deformation? *Geophys. J. Int.* **155**: 753–777.

Hedlin, M. A. H. and Shearer, P. M., 2000, An analysis of large-scale variations in small-scale mantle heterogeneity using global seismographic network recordings of precursors to PKP. *J. Geophys. Res.* **105**: 13655–13673.

Heirtzler, J. R., LePichon, X. and Baron, J. G., 1966, Magnetic anomalies over the Reykjannes Ridge. *Deep Sea Res.* **13**: 427–443.

Hellings, R. W., Adams, P. J., Anderson, J. D., Keesey, M. S., Lau, E. L. and Standish, E. M., 1983, Experimental test of the variability of *G* using Viking Lander ranging data. *Phys. Rev. Lett.* **51**: 1609–1612.

Helmholtz, H. von, 1856, On the interaction of natural forces. *Phil. Mag.* **11**: 489–578.

Henry, C. and Das, S., 2001, Aftershock zones of large shallow earthquakes: fault dimensions, aftershock area expansion and scaling relations. *Geophys. J. Int.* **147**: 272–293.

Hess, H. H, 1964, Seismic anisotropy of the upper mantle under oceans. *Nature* **203**: 629–631.

Hide, R., 1966, Free hydromagnetic oscillations of the Earth's core and the theory of the geomagnetic secular variation. *Phil. Trans. Roy. Soc. Lond.* **A259**: 615–650.

Hill, R., 1952, The elastic behaviour of a crystalline aggregate. *Proc. Phys. Soc.* **A65**: 349–354.

Hillgren, V. J. J., Schwager, B. and Boehler, R., 2005, Potassium as a heat source in the core? Metal–silicate partitioning of K and other metals. *Eos* (*Trans. Am. Geophys. Un.*) **86**(52), Fall meeting abstract MR13A-0086.

Hirao, N., Ohtani, E., Kondo, T., Endo, N., Kuba, T., Suzuki, T. and Kikegawa, T., 2006, Partitioning of potassium between iron and silicate at the core-mantle boundary. *Geophys. Res. Lett.* **33**: L08303. doi: 10:1029/2005GLO025324,2006.

Hollerbach, R. and Jones, C. A., 1995, On the magnetically stabilizing role of the Earth's inner core. *Phys. Earth Planet. Inter.* **87**: 171–181.

Holme, R. and deViron, O., 2005, Geomagnetic jerks and a high resolution length-of-day profile. *Geophys. J. Int.* **160**: 435–439.

Holmes, A., 1965, *Principles of Physical Geology*. London: Nelson.

Horton, B. K., 1999, Erosional control on the geometry and kinematics of the thrust belt development in the central Andes. *Tectonics* **18**(6): 1292–1304.

Hsu, W., Wasserburg, G. J. and Huss, G. R., 2000, High time resolution by use of the ^{26}Al chronometer in the multistage formation of a CAI. *Earth Plan. Sci. Lett.* **182**: 15–29.

Hurley, P. M., Hughes, H., Faure, G., Fairbairn, H. W. and Pinson, W. H., 1962, Radiogenic strontium-87 model of continent formation, *J. Geophys. Res.* **67**: 5315–5334.

Ide, S., Beroza, G. C., Prejean, S. G. and Ellsworth, W., 2003, Apparent break in earthquake scaling due to path and site effects on deep borehole recordings. *J. Geophys. Res.* **108**(B5): doi:10.1029/2001JB001617.

Isaak, D. G. and Masuda, K., 1995, Elastic and viscoelastic properties of α iron at high temperatures. *J. Geophys. Res.* **100**: 17689–17698.

Ishii, M., Shearer, P. M., Houston, H. and Vidale, J. E., 2005, Extent, duration and speed of the 2004 Sumatra–Andaman earthquake imaged by the Hi-Net array. *Nature* **435**: 933–936.

Ishikawa, Y. and Syono, Y., 1963, Order-disorder transformation and reverse thermoremanent magnetism in the $FeTiO_3$–Fe_2O_3 system. *J. Phys. Chem. Solids* **24**: 517–528.

Ita, J. and Stixrude, L., 1992, Petrology, elasticity, and composition of the mantle transition zone. *J. Geophys. Res.* **97**: 6849–6866.

Jackson, D. D., Shen, Z.-K., Potter, D., Ge, X.-B. and Sung, L., 1997, Southern California deformation. *Science* **277**: 1621–1622.

Jackson, I., Webb, S., Weston, L. and Boness, D., 2005, Frequency dependence of elastic wave speeds at high temperature: a direct experimental demonstration. *Phys. Earth Planet. Inter.* **148**: 85–96.

Jackson, J. A. and White, N. J., 1989, Normal faulting in the upper continental crust: observations from regions of active extension. *J. Struct. Geol.* **11**: 15–36.

Jaeger, J. C. and Cook, N. G., 1984, *Fundamentals of Rock Mechanics*, second edn. New York: Chapman and Hall.

James, D. E. (ed.), 1989, *The Encyclopedia of Solid Earth Geophysics*. New York: Van Nostrand-Reinhold.

Jeffreys, H., 1959, *The Earth, its Origin, History and Physical Constitution*, fourth edn. Cambridge: Cambridge University Press.

Johnston, M. J. S., Borcherdt, R. D., Linde, A. T. and Gladwin, M. T., 2006, Continuous borehole strain and pore pressure in the near field of the 28 September 2004 M6.0 Parkfield, California, earthquake: implications for nucleation, fault response, earthquake prediction and tremor. *Bull. Seism. Soc. Am.* **96**(4B): S56–S72.

Johnston, M. J. S. and Linde, A. T., 2002, Implications of crustal strain during conventional slow and silent earthquakes. In Lee, W., Kanamori, H., Jennings, P. and Kisslinger, C, *International Handbook of Earthquake and Engineering Seismology*, **81A**: 589–605. London: Academic Press.

Jones, L. E., Mori J. and Helmberger D. V., 1992, Short-period constraints on the upper mantle discontinuities *J. Geophys. Res.* **97**: 8765–8774.

Jones, G. M., 1977, Thermal interaction of the core and mantle and long term behaviour of the geomagnetic field. *J. Geophys. Res.* **82**: 1703–1709.

Kagan, Y. Y., 1991, Seismic moment distribution. *Geophys. J. Int.* **106**: 123–134.

Kagan Y. Y. 2002a, Seismic moment distribution revisited: I. Statistical results. *Geophys. J. Int.* **148**: 520–541.

Kagan Y. Y. 2002b, Seismic moment distribution revisited: II. Moment conservation principle. *Geophys. J. Int.* **149**: 731–754.

Kagan, Y. Y. and Jackson, D. D., 1994, Long-term probabilistic forecasting of earthquakes. *J. Geophys. Res.* **99**: 13685–13700.

Kagan, Y. Y. and Jackson, D. D., 2000, Probabilistic forecasting of earthquakes. *Int. J. Geophys.* **143**: 438–453.

Kagan, Y. Y. and Knopoff, L., 1987, Statistical short-term earthquake prediction. *Science* **236**: 1563–1567.

Kamo, S. L., Czamanske, G. K., Amelin, Y., Fedorenko, V. A., Davis, D. W. and Trofimov, V. R., 2003, Rapid eruption of Siberian flood-volcanic rocks and evidence for coincidence with the Permian–Triassic boundary and mass extinction at 251Ma. *Earth Plan. Sci. Lett.* **214**: 75–91.

Kanamori, H., 1977, The energy release in great earthquakes. *J. Geophys. Res.* **82**: 2981–2987.

Kanamori, H. and Anderson, D. L., 1975, Theoretical basis of some empirical relations in seismology. *Bull. Seism. Soc. Am.* **65**: 1073–1095.

Kanamori, H. and Brodsky, E. E., 2004, The physics of earthquakes. *Rep. Prog. Phys.* **67**: 1429–1496.

Kaneshima, S. and Helffrich, G., 1999, Dipping low-velocity layer in the mid-lower mantle: evidence for geochemical heterogeneity. *Science* **283**: 1888–1892.

Karato, S., 1993, Importance of anelasticity in the interpretation of seismic tomography. *Geophys. Res. Lett.* **20**: 1623–1626.

Kaufmann, G. and Lambeck, K., 2000, Mantle dynamics, postglacial rebound and the radial viscosity profile. *Phys. Earth Planet. Inter.* **121**: 301–324.

Kaula, W. M., 1968, *An Introduction to Planetary Physics: the Terrestrial Planets*. New York: Wiley.

Kawakatsu, H., 2006, Sharp and seismically transparent inner core boundary region revealed by an entire network observation of near vertical PKiKP. *Earth Planets Space* **58**(7): 855–863.

Keane, A., 1954, An investigation of finite strain in an isotropic material subjected to hydrostatic pressure and its seismological applications. *Australian J. Phys.* **7**: 322–333.

Keating, P. N., 1966, Effect of invariance requirements on the elastic strain energy of crystals with application to the diamond structure. *Phys. Rev.* **145**: 637–645.

Keen, C. E. and Barrett, D. L., 1971, A measurement of seismic anisotropy in the Northeast Pacific. *Can. J. Earth Sci.* **8**: 1056–1064.

Keilis-Borok, V., 2002, Earthquake prediction: state-of-the-art and emerging possibilities, *Ann. Rev. Earth Planet. Sci.* **30**: 1–33.

Keldysh, M. V., 1977, Venus exploration with Venera 9 and Venera 10 spacecraft. *Icarus* **30**: 605–625.

Kelvin, Lord (William Thomson), 1862, On the age of the Sun's heat. *Macmillan Mag.* **March 5, 1862**, 349–368.

Kelvin, Lord (William Thomson), 1863, On the secular cooling of the Earth. *Phil. Mag.* **25**: 1–14.

Kennett, B. L. N., 1983, *Seismic Wave Propagation in Stratified Media*. Cambridge: Cambridge University Press.

Kennett, B. L. N. and Engdahl, E. R., 1991, Traveltimes for global earthquake location and phase identification. *Geophys. J. Int.* **105**: 429–465.

Kennett, B. L. N., Engdahl, E. R. and Buland, A., 1995, Constraints on seismic velocities in the Earth from travel times. *Geophys. J. Int.* **122**: 108–124.

Kennett, B. L. N., Widiyantoro, S. and van der Hilst, R. D., 1998, Joint seismic tomography for bulk-sound and shear wavespeed in the Earth's mantle. *J. Geophys. Res.* **103**: 12469–12493.

Kent, D. V. and Smethurst, M. A., 1998, Shallow bias of magnetic inclinations in the Paleozoic and Precambrian. *Earth Plan. Sci. Lett.* **160**: 391–402.

Kesson, S. E. and Fitzgerald, J. D., 1992, Partitioning of MgO, FeO, NiO, MnO and Cr_2O_3 between magnesian silicate perovskite and magnesiowustite; implications for the origin of inclusions in diamonds and the composition of the lower mantle. *Earth Plan. Sci. Lett.* **111**: 229–240.

Kieffer, S. W., Getting, I. C. and Kennedy, G. C., 1976, Experimental determination of the thermal diffusivity of teflon, sodium chloride, quartz and silica. *J. Geophys. Res.* **81**: 3018–3024.

King, C., 1893, The age of the Earth. *Am. J. Science* **45**: 1–20.

King, S. D., 2002, Geoid and topography over subduction zones: the effect of phase transformations. *J. Gephys. Res.* **107** (B1). doi:10.1029/2000JB000141.

Kittel, C., 1949, Physical theory of ferromagnetic domains. *Rev. Mod. Phys.* **21**: 541–583.

Kittel, C., 1971, *Introduction to Solid State Physics*, fourth edn. New York: Wiley.

Kivelson, M. J. K., Khurana, K., Russell, C., Volwerk, M., Walker, R. J. and Zimmer, C., 2000, Galileo magnetometer measurements; a stronger case for a subsurface ocean at Europa. *Science* **289**: 1340–1343.

Knopoff, L., 1958, Energy release in earthquakes. *Geophys. J. Roy. Astr. Soc.* **1**: 44–52.

Knopoff, L., 1964, *Q. Revs. Geophys.* **2**: 625–660.

Knopoff, L., 2001, Rayleigh waves without cubic equations. *Computational Seismology* **32**: 31–37.

Kombayashi, T., Omori, S. and Maruyama, S., 2005, Experimental and theoretical study of dense hydrous magnesium silicates in the deep mantle. *Phys. Earth Planet. Inter.* **153**: 191–209.

Kong, X. and Bird, P., 1996, Neotectonics of Asia: thin shell finite-element with faults. In Yin, A. and Harrison, T. M. (eds.) *The Tectonic Evolution of Asia*. Cambridge: Cambridge University Press, pp. 18–34.

Kono, M. and Roberts, P. H., 2002, Recent geodynamo simulations and observations of the geomagnetic field. *Revs. Geophys.* **40**, doi: 10.1029/2000RG000102.

Konopliv, A. S. and Yoder, C. F., 1996, Venusian k_2 tidal Love number from Magellan and PVO tracking data. *Geophys. Res. Lett.* **23**: 1857–1860.

Kreemer, C., Holt, W. E. and Haines, A. J., 2003, An integrated model of present-day plate motions and plate boundary deformation. *Geophys. J. Int.* **154**: 8–34.

Kring, D. A. and Cohen, B. A., 2002, Cataclysmic bombardment throughout the inner Solar System 3.9–4.0 Ga. *J. Geophys. Res.* **107**(E2). doi 10.1029/2001JE001529.

Kuang, W. and Bloxham, J., 1997, An Earth-like numerical dynamo model. *Nature* **389**: 371–374.

Kyte, F. T., Smit, J. and Wasson, J. T., 1985, Siderophile interelement variations in the Cretaceous–Tertiary boundary sediments from Caravaca, Spain. *Earth Plan. Sci. Lett.* **73**: 183–195.

Lachenbruch, A. H., 1970, Crustal temperature and heat production: implications of the linear heat flow relation. *J. Geophys. Res.* **75**: 3291–3300.

Lachenbruch, A. H. and Sass, J. H., 1980, Heat flow and energetics of the San Andreas fault zone. *J. Geophys. Res.* **85**: 6185–6222 and **86**: 7171–7172.

Laj, C., Mazaud, A., Weeks, R., Fuller, M. and Herrero-Bervera, E., 1992, Statistical assessment of the preferred longitude bands for recent geomagnetic reversal records. *Geophys. Res. Lett.* **19**: 2003–2006.

Lamb, H., 1904, On the propagation of tremors over the surface of an elastic solid. *Phil. Trans. Roy. Soc. Lond.* **A203**: 1–42.

Lamb, S. and Davis, P., 2003, Cenozoic climate change as a possible cause for the rise of the Andes. *Nature* **425**: 792–797.

Lambeck, K., 1980, *The Earth's Variable Rotation*. Cambridge: Cambridge University Press.

Lambeck, K., 1990, Glacial rebound, sea level change and mantle viscosity. *Q. J. Roy. Astron. Soc.* **31**: 1–30.

Lambeck, K., Johnston, P., Smither, C. and Nakada, M., 1996, Glacial rebound of the British Isles – III. Constraints on mantle viscosity. *Geophys. J. Int.* **125**: 340–354.

Landau, L. D. and Lifshitz, E. M., 1975, *Theory of Elasticity*. Oxford: Pergamon Press.

Langel, R. A. and Estes, R. H., 1982, A geomagnetic field spectrum. *Geophys. Res. Lett.* **9**: 250–253.

Lapwood, E. R., 1949, The disturbance due to a line source in a semi-infinite elastic medium. *Phil. Trans. Roy. Soc., Lond.* **A242**: 63–100.

Larmor, J., 1919, How could a rotating body such as the Sun become a magnet? *Report of the 87th (1919) meeting* *of the British Association for the Advancement of Science*, pp. 159–160.

Laske, G. and Masters, G., 1998, Surface-wave polarization data and global anisotropic structure. *Geophys. J. Int.* **132**: 508–520.

Laske, G. and Masters, G., 2003, The Earth's free oscillations and the differential rotation of the inner core. In Dehant *et al.* (2003), pp. 5–21.

Lay, T. *et al.* (14 authors), 2005, The great Sumatra–Andaman earthquake of 26 December 2004. *Science* **308**: 1127–1133.

Lebedev, S., Chevrot, S. and van der Hilst, R. D., 2002, Seismic evidence for olivine phase changes at the 410- and 660-kilometer discontinuities. *Science* **296**: 1300–1302.

Lee, D. C., Halliday, A. N., Snyder, G. A. and Taylor, L. A., 1997, Age and origin of the moon. *Science* **278**: 1098–1103.

Lemoine, F. G. *et al.* (15 authors), 1998, *The Development of the Joint NASA GSFC and NIMA Geopotential Model EGM 96*, NASA Technical Paper, no. 1998–206861.

Lerch, F. J. *et al.* (20 authors), 1994, A geopotential model from satellite tracking, altimeter and surface gravity data: GEM-T3. *J. Geophys. Res.* **99**: 2815–2839.

Lin, J.-F., Jacobsen, S. D., Sturhahn, W., Jackson, J. M., Zhao, J. and Yoo, C.-S., 2006, Sound velocities of ferropericlase in the Earth's lower mantle. *Geophys. Res. Lett.* **33**: L22304, doi:10.1029/2006GL028099,2006.

Lister, J. R. and Buffett, B. A., 1995, The strength and efficiency of thermal and compositional convection in the geodynamo. *Phys. Earth Planet. Inter.* **91**: 17–30.

Liu, L.-G., 1976, Orthorhombic perovskite phase observed in olivine, pyroxene and garnet at high pressures and temperatures. *Phys. Earth Planet. Inter.* **11**: 289–298.

Long, C. and Christensen, N. I., 2000, Seismic anisotropy of South African upper mantle xenoliths. *Earth Plan. Sci. Lett.* **179**: 551–565.

Longuet-Higgins, M. S. and Ursell, F., 1948, Sea waves and microseisms. *Nature* **162**: 700.

Loper, D. E., 1978a, The gravitationally powered dynamo. *Geophys. J. R. Astron. Soc.* **54**: 389–404.

Loper, D. E., 1978b, Some thermal consequences of the gravitationally powered dynamo. *J. Geophys. Res.* **83**: 5961–5970.

Loper, D. E., 1984, The dynamical structures of D′′ and deep mantle plumes in a non-Newtonian mantle. *Phys. Earth Planet. Inter.* **33**: 56–67.

Loper, D. E., 1985, A simple model of whole mantle convection. *J. Geophys. Res.* **90**: 1809–1836.

Loper, D. E. and Stacey, F. D., 1983, The dynamical and thermal structure of deep mantle plumes. *Phys. Earth Planet. Inter.* **33**: 304–317.

Love, A. E. H., 1927, *A Treatise on the Mathematical Theory of Elasticity*, fourth edn. Cambridge: Cambridge University Press.

Lovell, A. C. B., 1954, *Meteor Astronomy*. Oxford: Clarendon Press.

Lowes, F. J., 1966, Mean values on sphere of spherical harmonic vector fields. *J. Geophys. Res.* **71**: 2179.

Lowes, F. J., and Wilkinson, I. 1963, Geomagnetic dynamo: a laboratory model. *Nature* **198**: 1158–1160.

Lowes, F. J., and Wilkinson, I. 1968, Geomagnetic dynamo: an improved laboratory model. *Nature* **219**: 717–718.

MacMillan, W. D., 1958, *Theory of the Potential*. New York: Dover (reprinted from 1930 edition).

Macouin, M., Valet, G. P. and Besse, J., 2004, Long-term evolution of the geomagnetic dipole moment. *Phys. Earth Planet. Inter.* **147**: 239–246.

Madariaga, R., 1976, Dynamics of an expanding circular fault. *Bull. Seism. Soc. Am.* **66**: 639–667.

Maggi, A., Debayle, E., Priestley, K. and Barruol, G., 2006, Multimode surface waveform tomography of the Pacific Ocean: a closer look at the lithospheric cooling signature. *Geophys. J. Int.* **166**: 1384–1397.

Malkus, W. V. R., 1963, Precessional torques as the cause of geomagnetism. *J. Geophys. Res.* **68**: 2871–2886.

Malkus, W. V. R., 1989, An experimental study of global instabilities due to tidal (elliptical) distortion of a rotating elastic cylinder. *Geophys. Astrophys. Fluid. Dyn.* **48**: 123–134.

Manga, M. and Jeanloz, R., 1997, Thermal conductivity of corundum and periclase and implications for the lower mantle. *J. Geophys. Res.* **102**: 2999–3008.

Mansinha, L. and Smylie, D. E., 1971, The displacement fields of inclined faults. *Bull. Seism. Soc. Am.* **61**: 1433–1440.

Mao, W. L., Mao, H.-K., Sturhahn, W., Zhao, J., Prakapenka, V. B., Meng, Y., Shu, J., Fei, Y. and Hemley, R. J., 2006, Iron-rich postperovskite and the origin of ultralow-velocity zones. *Science* **312**: 564–565.

Margot, J.-L., Peale, S. J., Jurgens, R. F., Slade, M. A. and Holin, I. V., 2007, Large longitude libration of Mercury reveals a molten core. *Science* **316**: 710–714.

Marone, C. J., Scholz, C. H. and Bilham, R., 1991, On the mechanics of earthquake afterslip. *J. Geophys. Res.* **96**(5): 8441–8452.

Marquering, H., Dahlen, F. A., and Nolet, G., 1999, Three-dimensional sensitivity kernels for finite-frequency traveltimes: the banana doughnut paradox. *Geophys. J. Int.* **137**: 805–815.

Masters, G. and Gilbert, F., 1981, Structure of the inner core inferred from observations of its spheroidal shear modes. *Geophys. Res. Lett.* **8**: 569–571.

Masters, G. and Gubbins, D., 2003, On the resolution of density within the Earth. *Phys. Earth Planet. Inter.* **140**: 159–167.

Masters, G., Laske, G., Bolton, H. and Dziewonski, A. M., 2000, The relative behaviour of shear velocity, bulk sound speed, and compressional velocity in the mantle: implications for chemical and thermal structure. In Karato, S.-I. *et al*, eds., *Earth's deep interior: mineral physics and tomography from the atomic to the global scale*. Geophysical Monograph Series **117**: 63–87. Washington: American Geophysical Union.

Masters, T. G. and Widmer, R., 1995, Free oscillations: frequencies and attenuation. In Ahrens (1995a), pp. 104–125.

Mathews, P. M., Buffett, B. A. and Shapiro, I. I., 1995, Love numbers for diurnal tides: relation to wobble admittances and resonance expansions. *J. Geophys. Res.* **100**: 9935–9948.

Mathews, P. M., Herring, T. A. and Buffett, B. A., 2002, Modeling nutation and precession: new nutation series for nonrigid Earth and insights into the Earth's interior. *J. Geophys. Res.* **107**(B4). 10.1029/2001JB000390.2002.

Maxwell, A. E., Von Herzen, R. P., Hsü, K. J., Andrews, J. E., Saito, T., Percival, S., Milow, E. D. and Boyce, R. E., 1970, Deep sea drilling in the South Atlantic. *Science* **168**: 1047–1059.

McArdle, N. J., Halls, H. C. and Shaw, J., 2004, Rock magnetic studies and a comparison between microwave and Thellier paleointensities for Canadian Precambrian dykes. *Phys. Earth Planet. Inter.* **147**: 247–254.

McDonough, W. F. and Sun, S.-S., 1995, The composition of the Earth. *Chem. Geology* **120**: 223–253.

McDougall, I., 1981, ^{40}Ar/^{39}Ar age spectra for the KBS tuff, Koobi Fora formation. *Nature* **294**: 120–124.

McDougall, I. and Harrison, T. M., 1999, *Geochronology and Thermochronology by the ^{40}Ar/^{39}Ar Method*. New York: Oxford University Press.

McDougall, I., Maier, R., Sutherland-Hawkes, P. and Gleadow, A. J. W., 1980, K-Ar age estimate for the KBS tuff, East Turkana, Kenya. *Nature* **284**: 230–234.

McDougall, I. and Tarling, D. H., 1963, Dating of polarity zones in the Hawaiian islands. *Nature* **200**: 54–56.

McFadden, P. L. and Merrill, R. T., 1984, Lower mantle convection and geomagnetism. *J. Geophys. Res.* **89**: 3354–3362.

McFadden, P. L. and Merrill, R. T., 1995, History of the Earth's magnetic field and possible connections to core–mantle boundary processes. *J. Geophys. Res.* **100**: 307–316.

McFadden, P. L., Merrill, R. T. and McElhinny, M. W., 1988, Dipole/quadrupole modelling of paleosecular variation. *J. Geophys. Res.* **93**: 11583–11588.

McFadden, P. L., Merrill, R. T., McElhinny, M. W. and Lee, S., 1991, Reversals of the Earth's magnetic field and temporal variations of the dynamo families. *J. Geophys. Res.* **96**: 3923–3933.

McGarr, A., 1999, On relating apparent stress to the stress causing earthquake fault slip. *J. Geophys. Res.* **104**(B2): 3003–3011.

McKenzie, D., Jackson, J. and Priestley, K., 2005, Thermal structure of oceanic and continental lithosphere. *Earth Plan. Sci. Lett.* **233**: 337–349.

McLennan, S. M., 1995, Sediments and soils: chemistry and abundances. In Ahrens (1995c). pp. 8–19.

McNutt, M. K., 1998, Superswells. *Revs. Geophys.* **36**: 211–244.

McQueen, R. G. and Marsh, S. P, 1966, Shock wave compression of iron-nickel alloys and the Earth's core. *J. Geophys. Res.* **71**: 1751–1756.

McQueen, R. G., Marsh, S. P. and Fritz, J. N., 1967, Hugoniot equation of state of twelve rocks. *J. Geophys. Res.* **72**: 4999–5036.

McSween, H. Y., 1999, *Meteorites and their Parent Planets*. Cambridge: Cambridge University Press.

Mei, S. and Kohlstedt, D. L., 2000a, Influence of water on plastic deformation of olivine aggregates 1: Diffusion creep regime. *J. Geophys. Res.* **105**: 21457–21469.

Mei, S. and Kohlstedt, D. L., 2000b, Influence of water on plastic deformation of olivine aggregates 2: Dislocation creep regime. *J. Geophys. Res.* **105**: 21471–21481.

Meredith, P. G. and Atkinson, B. K., 1983, Stress corrosion and acoustic emission during tensile crack propagation in Whin Sill dolerite and other basic rocks. *Geophys. J. Roy. Astr. Soc.* **75**: 1–21.

Merrill, R. T., McElhinny, M. W. and McFadden, P. L., 1996, *The Magnetic Field of the Earth: Paleomagnetism, the Core and the Deep Mantle*. San Diego: Academic Press.

Merrill, R. T. and McFadden, P. L., 1999, Geomagnetic polarity transitions. *Rev. Geophys.* **37**: 201–226.

Mitrovica, J. X., 1996, Haskell [1935] revisited. *J. Geophys. Res.* **101**: 555–569.

Mitrovica, J. X. and Forte, A. M., 1997, Radial profile of mantle viscosity: Results from the joint inversion of convection and postglacial rebound observable. *J. Geophys. Res.* **102**: 2751–2769.

Mitrovica, J. X. and Peltier, W. R., 1993, Present day secular variations in the zonal harmonics of Earth's geopotential. *J. Geophys. Res.* **98**: 4509–4526.

Mogi, K., 1958, Relations between the eruptions of various volcanoes and the deformation of the ground surface around them. *Bull. Earthq. Res. Inst. Univ. Tokyo* **36**: 99–134.

Molnar, P. and Atwater, T., 1973, Relative motion of hotspots in the mantle. *Nature* **246**: 288–291.

Montagner, J.-P., Griot-Pommera, D.-A. and Lave, J., 2000, How to relate body wave and surface wave anisotropy? *J. Geophys. Res.* **105**: 19015–19027.

Montagner, J.-P. and Kennett, B. L. N., 1996, How to reconcile body wave and normal mode reference models. *Geophys. J. Int.* **125**: 229–248.

Montagner, J.-P. and Tanimoto, T., 1991, Global upper mantle tomography of seismic velocities and anisotropies. *J. Geophys. Res.* **96**: 20337–20351.

Montelli, R., Nolet, G., Dahlen, F. A., Masters, G., Engdahl, E. R. and Hung, S.-H., 2004, Finite-frequency tomography reveals a variety of plumes in the mantle. *Science* **30**: 338–343.

Morgan, W. J., 1971, Convection plumes in the lower mantle. *Nature* **230**: 42–43.

Morozov, I. B. and Smithson, S. B., 2000, Coda of long-range arrivals from nuclear explosions. *Bull. Seism. Soc. Am.* **90**: 929–939.

Morris, J. D., Leeman, W. P. and Tera, F., 1990, The subducted component in island arc lavas: constraint from Be isotopes and B-Be systematics. *Nature* **344**: 31–36.

Mukhopadhyay, S. and Nittler, L., 2004, Report in Yearbook 02/03, p. 69. Washington: Carnegie Institution.

Murakami, M., Hirose, K., Kawamura, K., Sata, N. and Ohishi, Y., 2004, Post-perovskite phase transition in $MgSiO_3$. *Science* **304**: 855–858.

Murthy, V. R., van Westrenen, W. and Fei, Y., 2003, Experimental evidence that potassium is a substantial radioactive heat source in planetary cores. *Nature* **423**: 163–165.

Nadeau, R. M. and Dolenc, D., 2005, Nonvolcanic tremors deep beneath the San Andreas Fault. *Science* **307**: 389–390.

Nagata, T., 1953, *Rock Magnetism*, first edn. Tokyo: Maruzen.

Nagata, T., 1979, Meteorite magnetism and the early solar system magnetic field. *Phys. Earth Planet. Inter.* **20**: 324–341.

Nakiboglu, S. M., 1982, Hydrostatic theory of the Earth and its mechanical implications. *Phys. Earth Planet. Inter.* **28**: 302–311.

Narayan, C. and Goldstein, J. I., 1985, A major revision of iron meteorite cooling rates – an experimental study of the growth of the Widmanstätten pattern. *Geochim. Cosmochim. Acta* **49**: 397–410.

Navon, O. and Wasserburg, G. J., 1985, Self-shielding in O_2 – a possible explanation of oxygen isotope anomalies in meteorites. *Earth Plan. Sci. Lett.* **73**: 1–16.

Nawa, K., Sudo, N., Fukao, Y., Sato., T., Aoyama, Y. and Shibuya, K., 1998, Incessant excitation of the Earth's free oscillations. *Earth Space Sci.* **50**: 3–8.

Néel, L. 1955, Some theoretical aspects of rock magnetism. *Adv. Phys.* **4**: 191–243.

Ness, N. F., 1994, Intrinsic magnetic fields of the planets: Mercury to Neptune. *Phil. Trans. Roy. Soc. Lond.* **A349**: 249–260.

Newsom, H. E., 1995, Composition of the solar system, planets, meteorites and major terrestrial reservoirs. In Ahrens (1995a), pp. 159–189.

Nieto, M. M., 1972, *The Titius–Bode Law of Interplanetary Distances: its History and Theory*. Oxford: Pergamon.

Nimmo, F., Price, G. D., Brodholt, J. and Gubbins, D., 2004, The influence of potassium on core and geodynamo evolution. *Geophys. J. Int.* **156**: 363–376.

Nishimura, C. E. and Forsyth, D. W., 1989, The anisotropic structure of the upper mantle in the Pacific. *Geophys. J.* **96**: 203–229.

Nittler, L., 2003, Presolar stardust in meteorites: recent advances and scientific frontiers. *Earth Plan. Sci. Lett.* **209**: 259–273.

Norton, I. O., 1995, Plate motions in the north Pacific: the 43 Ma nonevent. *Tectonics* **14**(5): 1080–1094.

Nyblade, A. A. and Robinson, S. W., 1994, The African superswell. *Geophys. Res. Lett.* **21**: 765–768.

Oganov, A. R., Brodholt, J. P. and Price, G. D., 2000, Comparative study of quasiharmonic lattice dynamics, molecular dynamics and Debye model applied to $MgSiO_3$ perovskite. *Phys. Earth Planet. Inter.* **122**: 277–288.

Ogata, Y., 1998, Space-time point process models for earthquake occurrences. *Annals Inst. Statistical Mechanics* **50**: 379–402.

Ogino, K., Nishiwacki, A. and Hosotani, Y., 1984, Density of molten Fe-C alloys. *J. Japan Inst. Metals* **48**: 1004–1010.

Ohtani, E., Litasov, K., Suzuki, A. and Kondo, T., 2001, Stability field of a new hydrous phase, δ-AlOOH, with implications for water transport in the deep mantle. *Geophys. Res. Lett.* **28**: 3991–3993.

Okada, Y., 1985, Surface deformation due to shear and tensile faults in a half space. *Bull. Seism. Soc. Am.* **75**: 1135–1154.

Okal, E. A., 2001, 'Detached' deep earthquakes: are they really? *Phys. Earth Planet. Inter.* **127**: 109–143.

Okuchi, T., 1997, Hydrogen partitioning into molten iron at high pressure: implications for the Earth's core. *Science* **278**: 1781–1784.

Okuchi, T., 1998, The melting temperature of iron hydride at high pressures and its implication for the temperature of the Earth's core. *J. Phys. Condensed Matter* **10**: 11595–11598.

Oliver, J., 1962, A summary of observed seismic wave dispersion. *Bull. Seism. Soc. Am.* **52**: 81–86.

Olsen, N., 2002, A model of the geomagnetic field and its secular variation for the epoch 2000 estimated from Ørsted data. *Geophys. J. Int.* **149**: 454–462.

Olsen, P. E. *et al.* (10 authors), 2002, Ascent of dinosaurs linked to an iridium anomaly at the Triassic–Jurassic boundary. *Science* **296**: 1305–1307.

Olson, P., 1983, Geomagnetic polarity reversals in a turbulent core. *Phys. Earth Planet. Inter.* **33**: 260–274.

Olson, P. and Aurnou, J. 1999, A polar vortex in the Earth's core. *Nature* **402**: 170–173.

Omori, F. J., 1894, On after-shocks of earthquakes. *College of Science, Imperial University of Tokyo* **7**: 111–200.

Opdyke, N. D. and Channell, J. E. T., 1996, *Magnetic Stratigraphy*. San Diego: Academic Press.

Opdyke, N. D., Kent, D. V. and Lowrie, W., 1973, Details of magnetic polarity transitions recorded in a high deposition rate deep sea core. *Earth Plan. Sci. Lett.* **20**: 315–324.

Oversby, V. M. and Ringwood, A. E., 1971, Time of formation of the Earth's core. *Nature* **234**: 463–465.

Ozima, M. and Podosek, F. A., 1999, Formation age of Earth from ^{129}I/^{127}I and ^{244}Pu/^{238}U systematics and the missing Xe. *J. Geophys. Res.* **104**: 25493–25499.

Padhy, S, 2005, A scattering model for seismic attenuation and its global applications. *Phys. Earth Planet. Int.* **148**: 1–12.

Pagiatakis, S. D., Yin, H. and El-Gelil, M. A., 2007, Least squares self-coherency analysis of superconducting gravimeter records in search for the Slichter triplet. *Phys. Earth Planet. Inter.* **160**: 108–123.

Panning, M. P. and Romanowicz, B. A., 2006, A three dimensional radially anisotropic model of shear velocity in the whole mantle. *Geophys. J. Int.* **167**: 361–379.

Parkinson, W. D., 1983, *Introduction to Geomagnetism*. Edinburgh: Scottish Academic Press.

Paterson, M. S. and Weiss, L. E., 1961, Symmetry concepts in the structural analysis of deformed rocks. *Geol. Soc. Am. Bull.* **72**: 841–882.

Peale, S. J., Cassen, P. and Reynolds, R. P., 1979, Melting of Io by tidal dissipation. *Science* **203**: 892–894.

Pearce, S. J. and Russell, R. D., 1990, Inversion of cosmogenic nuclide data from iron meteorites. *Canad. J. Earth Sci.* **68**: 1312–1321.

Peltier, W. R., 1982, Dynamics of the ice age Earth. *Adv. Geophys.* **24**: 1–146.

Peltier, W. R., 1998, Postglacial variations in the level of the sea: implications for climate dynamics. *Rev. Geophys.* **36**: 603–689.

Peltier, W. R., 2004, Global glacial isostasy and the surface of the ice age Earth: the Ice-5 g (Vm2) model and Grace. *Ann. Rev. Earth Plan. Sci.* **32**: 111–149.

Peltzer, G., Crampé, F. and King, G., 1999, Evidence of nonlinear elasticity in the crust from the Mw7.6 Manyi (Tibet) earthquake. *Science* **286**: 272–276.

Pesonen, L. J., Elming, S.-A., Mertanen, S., Pisarevsky, S., D'Agrella-Filho, M. S., Meert, J. G., Schmidt, P. W., Abrahamsen, N. and Bylund, G., 2003, Palaeomagnetic configuration of the continents during the Proterozoic. *Tectonophysics* **375**: 289–324.

Plafker, G., 1965, Tectonic deformation associated with the 1964 Alaska earthquake. *Science* **148**: 1675–1687.

Poirier, J. -P., 1988, Transport properties of liquid metals and viscosity of the Earth's core. *Geophys. J. R. Astron. Soc.* **92**: 99–105.

Poirier, J. -P., 1994, Light elements in the Earth's core: a critical review. *Phys. Earth Planet. Inter.* **85**: 319–337.

Poirier, J.-P., 2000, *Introduction to the Physics of the Earth's Interior*, second edn. Cambridge: Cambridge University Press.

Poirier, J.-P. and Tarantola, A., 1998, A logarithmic equation of state. *Phys. Earth Planet. Inter.* **109**: 1–8.

Pollack, H. N. and Huang, S., 2000, Climate reconstruction from subsurface temperatures. *Ann. Rev. Earth Plan. Sci.* **28**: 339–365.

Pollack, H. N., Hurter, S. J. and Johnson, J. R., 1993, Heat flow from the Earth's interior: analysis of the global data set. *Rev. Geophys.* **31**: 267–280.

Poupinet, G. R., Pillet, R. and Souriau, A., 1983, Possible heterogeneity of the Earth's core deduced from PKIKP travel times. *Nature* **305**: 204–206.

Prentice, A. J. R., 1986, Uranus: predicted origin and composition of its atmosphere, moons and rings. *Phys. Lett.* **A114**: 211–216.

Prentice, A. J. R., 1989, Neptune: predicted origin and composition of a regular satellite system. *Phys. Lett.* **A140**: 265–270.

Proudman, J., 1953, *Dynamical Oceanography*. London: Methuen.

Rabinowicz, E., 1965, *Friction and Wear of Materials*. New York: Wiley.

Rädler, K.-H. and Cēbers, A. (Eds.), 2002, MHD dynamo experiments. *Magnetohydrodynamics* **38**: 3–217.

Raitt, R. W., Shor, G. G., Francis, T. G. J. and Morris G. B., 1969, Anisotropy of the Pacific upper mantle. *J. Geophys. Res.* **74**: 3095–3109.

Rapp, R. H. and Pavlis, N. K., 1990, The development and analysis of geopotential coefficient models to spherical harmonic degree 360. *J. Geophys. Res.* **95**: 21885–21911.

Ray, R. D., Eanes, R. J. and LeMoine, F. G., 2001, Constraints on energy dissipation in the Earth's body tide from satellite tracking and altimetry. *Geophys. J. Int.* **144**: 471–480.

Reasenberg, P. A., 1999, Foreshock occurrence before large earthquakes, *J. Geophys. Res.* **104**(B3): 4755–4768.

Reid, H. F., 1910, *The California Earthquake of April 18, 1906. II. The Mechanics of the Earthquake*. Washington: Carnegie Institution.

Reinecker, J., Heidbach, O., Tingay, M., Connolly, P. and Müller, B., 2004, The 2004 release of *The World Stress Map*. (www.world-stress-map.org).

Rhie, J. and Romanowicz, B., 2004, Excitation of the Earth's free oscillations by atmosphere–ocean–seafloor coupling. *Nature* **431**: 552–556.

Richards, M. A. and Engebretson, D. C., 1992, Large scale mantle convection and the history of subduction. *Nature* **355**: 437–440.

Richardson, R. M., 1992, Ridge forces, absolute plate motions and the intraplate stress field. *J. Geophys. Res.* **97**(8): 11739–11748.

Richter, C. F., 1958, *Elementary Seismology*. San Francisco: Freeman.

Rigden, S. M., Gwanmesia, G. D., Fitzgerald, J. D., Jackson, I. and Liebermann, R. C., 1991, Spinel elasticity and seismic structure of the transition zone of the mantle. *Nature* **34**: 143–145.

Rikitake, T., 1966, *Electromagnetism and the Earth's Interior*. Amsterdam: Elsevier.

Ringwood, A. E., 1966, Chemical evolution of the terrestrial planets. *Geochim. Cosmochim. Acta* **30**: 41–104.

Ringwood, A. E., 1989, Flaws in the giant impact hypothesis of lunar origin. *Earth Plan. Sci. Lett.* **95**: 208–214.

Ritsema, J., van Heijst, H. J. and Woodhouse, J. H., 1999, Complex shear velocity structure imaged beneath Africa and Iceland. *Science* **286**: 1925–1928.

Roberts, P., 1987, Origin of the main field: dynamics. In Jacobs, J. A. (ed.), *Geomagnetism* Vol. 2. London: Academic Press, pp. 251–306.

Roberts, P. H. and Gubbins, D., 1987, Origin of the main field: dynamics. In Jacobs, J. A. (ed.), *Geomagnetism* Vol. 2. London: Academic Press, pp. 185–249.

Robertson, G. S. and Woodhouse, J. H., 1996a, Ratio of relative *S* to *P* heterogeneity in the lower mantle. *J. Geophys. Res.* **101**: 20041–20052.

Robertson, G. S. and Woodhouse, J. H., 1996b, Constraints on lower mantle physical properties from seismology and mineral physics. *Earth Planet. Sci. Lett.* **143**: 197–205.

Robock, A., 2000, Volcanic eruptions and climate. *Rev. Geophys.* **38**: 191–219.

Robock, A., 2003, Introduction: Mount Pinatubo as a test of climate feedback mechanisms. In Robock, A. and Oppenheimer, C. (eds.) *Volcanism and the Earth's Atmosphere*. Washington: American Geophysical Union, pp. 1–8.

Rogers, G. and Dragert, H., 2003, Episodic tremor and slip on the Cascadia subduction zone: the chatter of silent slip. *Science* **300**: 1942–1943.

Roth, M., Müller, G. and Snieder, R., 1993, Velocity shifts in random media. *Geophys. J. Int.* **115**: 552–563.

Runnegar, B., 1982, The Cambrian explosion: animals or fossils? *J. Geol. Soc. Australia* **29**: 395–411.

Rutherford, E. and Soddy, F., 1903, Radioactive change. *Phil. Mag.* (*Series 6*) **5**: 1576–1591.

Ryder, G., 1990, Lunar samples, lunar accretion and the early bombardment of the Moon. *EOS (Trans. AGU Spring Meeting Supplement)* **71**: 313 and 322–323 (March 6, 1990).

Ryder, G. and Mojzsis, S. J., 1998, Accretion to the Earth and Moon around 3.85 Ga: what is the evidence? *EOS (Trans. AGU Fall Meeting Supplement)* **79**(45): F48 (Abstract U22B-10).

Sanloup, C., Guyot, F., Gillet, P., Fiquet, G., Hemley, R. J., Mezouar, M. and Martinez, I., 2000, Structural changes in liquid Fe at high pressures and high temperatures from synchrotron X-ray diffraction. *Europhys. Lett.* **52**: 151–157.

Sasatani, T., 1989, Deep earthquakes. In James (1989), pp. 174–181.

Schneider, J. F. and Sacks, I. S., 1992, Subduction of the Nazca plate beneath central Peru from local earthquakes. Unpublished manuscript.

Scholz, C. H., 1990, *The Mechanics of Earthquakes and Faulting*. Cambridge: Cambridge University Press.

Schubert, G., Masters, G., Olson P. and Tackley, P., 2004, Superplumes or plume clusters? *Phys. Earth Planet. Int.* **146**: 147–162.

Secco, R. A., 1995, Viscosity of the outer core. In Ahrens (1995b), pp. 218–226.

Sella, G. F., Stein, S., Dixon, T. H., Craymer, M., James, T. S., Mazzotti, S. and Dokka, R. K., 2007, Observation of glacial isostatic adjustment in "stable" North America with GPS. *Geophys. Res. Lett.* **34**: L02306. doi:10.1029/2006GL027081.

Shaw, B. E., 1993, Generalized Omori law for aftershocks and foreshocks from simple dynamics. *Geophys. Res. Letters* **20**: 907–910.

Shaw, J., 1974, A new method of determining the magnitude of the palaeomagnetic field. *Geophys. J. R. Astron. Soc.* **39**: 133–141.

Shaw, J. and Sherwood, G., 1991, Palaeointensity and reversal frequency – are they related? *Geophy. Astrophy. Fluid Dyn.* **60**: 135–140.

Shearer, P. M., 1990, Seismic imaging of upper mantle structure with new evidence for a 520 km discontinuity. *Nature* **344**: 121–126.

Sheriff, R. E., and Geldart, L. P., 1982, *Exploration Seismology, Vol. 1: History, Theory and Data Acquisition*. Cambridge: Cambridge University Press.

Sibson, R. H. and Xie, G., 1998, Dip range for intracontinental reverse fault ruptures: truth not stranger than friction. *Bull. Seism. Soc. Am.* **88**: 1014–1022.

Silver, P. G., 1996, Seismic anisotropy beneath the continents: probing the depths of geology. *Ann. Rev. Earth Planet. Sci.* **24**: 385–432.

Singh, S. K. and Ordaz, M., 1994, Seismic energy release in Mexican subduction zone earthquakes. *Bull. Seism. Soc. Am.* **84**: 1533–1550.

Slater, J. C., 1939, *Introduction to Chemical Physics*. New York: McGraw-Hill.

Sleep, H. N., 1990, Hot spots and mantle plumes: some phenomenology. *J. Geophys. Res.* **95**: 6715–6736.

Slichter, L. B., 1967, Spherical oscillations of the earth, *Geophys. J. R. Astron. Soc.* **14**: 171–177.

Smith, S. W., 1967, Free vibrations of the Earth. In Runcorn, S. K. (ed.) *International Dictionary of Geophysics* (2 vols.) Oxford: Pergamon, pp. 344–346.

Smyth, J. R., and McCormick, T. C., 1995, Crystallographic data for minerals. In Ahrens, T. J. (1995b), pp. 1–17.

Sneddon, I. N., 1980, *Special Functions of Mathematical Physics and Chemistry*, third edn. Edinburgh: Oliver and Boyd.

Solheim, L. P. and Peltier, W. R., 1994, Phase boundary deflections at 660 km depth and episodically layered isochemical convection in the mantle. *J. Geophys. Res.* **99**: 15861–15875.

Solomatov, V. S. and Stevenson, D. J., 1994, Can sharp seismic discontinuities be caused by non-equilibrium phase transitions? *Earth Plan. Sci. Lett.* **125**: 267–279.

Song, X. and Richards, P. G., 1996, Seismological evidence for differential rotation of the Earth's inner core. *Nature* **382**: 221–224.

Souriau, A., Roudil, P. and Moynot, B. 1997, Inner core differential rotation: facts and artifacts. *Geophys. Res. Lett.* **24**: 2103–2106.

Spetzler, J. and Snieder, R., 2004, Tutorial, the Fresnel volume and transmitted waves. *Geophysics* **69**: 653–663.

Stacey, F. D., 1973, The coupling of the core to the precession of the Earth. *Geophys. J. R. Astron Soc.* **33**: 47–55.

Stacey, F. D., 2000, Kelvin's age of the Earth paradox revisited. *J. Geophys. Res.* **105**: 13155–13158.

Stacey, F. D., 2005, High pressure equations of state and planetary interiors. *Reps. Prog. Phys.* **68**: 341–383.

Stacey, F. D. and Anderson, O. L., 2001, Electrical and thermal conductivities of Fe–Ni–Si alloy under core conditions. *Phys. Earth Planet. Inter.* **124**: 153–162.

Stacey, F. D. and Davis, P. M., 2004, High pressure equations of state with applications to the lower mantle and core. *Phys. Earth Planet. Inter.* **142**: 137–184.

Stacey, F. D. and Irvine, R. D., 1977, A simple dislocation theory of melting. *Australian J. Phys.* **30**: 641–646.

Stacey, F. D. and Isaak, D. G., 2003, Anharmonicity in mineral physics: a physical interpretation. *J. Geophys. Res.* **108(B9)**: 2440. doi:10.1029/2002JB002316,2003.

Stacey, F. D. and Loper, D. E., 1983, The thermal boundary layer interpretation of D″ and its role as a plume source. *Phys. Earth Planet. Inter.* **33**: 45–55.

Stacey, F. D. and Loper, D. E., 1984, Thermal histories of the core and mantle. *Phys. Earth Planet. Inter.* **36**: 99–115.

Stacey, F. D. and Loper, D. E., 2007, A revised estimate of the conductivity of iron alloy at high pressure and implications for the core energy balance. *Phys. Earth Planet. Inter.* **161**: 13–18.

Stacey, F. D., Spiliopoulos, S. S. and Barton, M. A., 1989, a critical re-examination of the thermodynamic basis of Lindemann's melting law. *Phys. Earth Planet. Inter.* **55**: 201–207.

Stacey, F. D. and Stacey, C. H. B., 1999, Gravitational energy of core evolution: implications for thermal history and geodynamo power. *Phys. Earth Planet. Inter.* **110**: 83–93.

Stanley, S., Bloxham. J., Hutchison, W. E. and Zuber, M. T., 2005, Thin shell dynamo models consistent with Mercury's weak observed magnetic field. *Earth Planet. Sci. Lett.* **234**: 27–38.

Stein, C. A., 1995, Heat flow from the Earth. In Ahrens (1995a), pp. 144–158.

Stein, C. A. and Stein, S., 1992, A model for the global variation in oceanic depth and heat flow with lithospheric age. *Nature* **359**: 123–129.

Stein, C. A. and Stein, S., 1994, Constraints on hydrothermal heat flux through oceanic lithosphere from global heat flux. *J. Geophys. Res.* **99**: 3081–3095.

Stein, R. S., 1999, The role of stress transfer in earthquake occurrence. *Nature* **402**: 605–609.

Stein, S. and Wysession, M., 2003, *An Introduction to Seismology, Earthquakes and Earth Structure*. Oxford: Blackwell.

Stephenson, F. R. and Morrison, L. V., 1995, Long-term fluctuations in the Earth's rotation. *Phil. Trans. Roy. Soc. Lond.* **A351**: 165–202.

Stevenson, D. J., 2003, Planetary magnetic fields. *Earth Plan. Sci. Lett.* **208**: 1–11.

Stevenson, D., 2005, Earthquakes and tsunamis: what physics is interesting? *Physics Today* **June 2005**: 10–11.

Stoneley, R., 1924, Elastic waves at the surface of separation of two solids. *Proc. Roy. Soc. Lond.* **A106**: 416–420.

Strutt, R. J., 1906, On the distribution of radium in the Earth's crust and on the Earth's internal heat. *Proc. Roy. Soc. Lond.* **A77**: 472–485.

Sturhahn, W., Jackson, J. M. and Lin, J.-F., 2005, The spin state of iron in minerals of the Earth's lower mantle. *Geophys. Res. Lett.* **32**: L12307, doi:10.1029/2005GL022802,2005.

Su, W. J. and Dziewonski, A. M., 1997, Simultaneous inversion for 3D variations in shear and bulk velocity in the mantle. *Phys. Earth Planet. Inter.* **100**: 135–156.

Su, W. J., Dziewonski, A. M and Jeanloz, R., 1996, Planet within a planet: rotation of the inner core of the Earth. *Science* **274**: 1883–1887.

Sumita, I. and Yoshida, S., 2003, Thermal interactions between the mantle, outer and inner cores, and the resulting structural evolution of the core. In Dehant *et al.* (2003), pp. 213–231.

Tackley, P. J., Stevenson, D. J., Glatzmaier, G. A. and Schubert, G., 1994, Effects of multiple phase transitions in a three-dimensional spherical model of convection in the Earth's mantle. *J. Geophys. Res.* **99**: 15877–15901.

Takahashi, F., Matsushima, M. and Honkura, Y., 2005, Simulations of a quasi-Taylor state geomagnetic field including polarity reversals on the Earth simulator. *Science* **309**: 459–461.

Tarling, D., 1971, *Principles and Applications of Palaeomagnetism*. London: Chapman and Hall.

Tarling, D. H., 1989, Archaeomagnetism. In James (1989), pp. 33–37.

Tatsumoto, M., 1966, Genetic relationships of ocean basalts as indicated by lead isotopes. *Science* **153**: 1094–1101.

Tatsumoto, M., Knight, R. J. and Allègre, C. J., 1973, Time differences in the formation of meteorites as determined by the ratio of lead-207 to lead-206. *Science* **180**: 1279–1283.

Tauxe, L., 2006, Long-term trends in paleointensity: the contribution of DSDP/ODP submarine basalt glass collections. *Phys. Earth Planet. Inter.* **156**: 223–241.

Tera, F., 2003, A lead isotope method for the accurate dating of disturbed geological systems: numerical demonstrations, some applications and implications. *Geochim. Cosmochim. Acta* **67**: 3687–3715.

Tera, F., Papanastassiou, D. A. and Wasserburg, G. J., 1974, Isotopic evidence for a terminal lunar cataclysm. *Earth Plan. Sci. Lett.* **22**: 1–21.

Thatcher, W., 1983, Nonlinear strain buildup and the earthquake cycle on the San Andreas fault. *J. Geophys. Res.* **88**: 5893–5902.

Thellier, E. and Thellier, O., 1959, Sur l'intensité du champ magnétique terrestre dans le passé historique et géologique. *Ann. Geophys.* **15**: 285–376.

Tilton, G. R. and Steiger, R. H., 1965, Lead isotopes and the age of the Earth. *Science* **150**: 1805–1808.

Toon, O. B., Zahnle, K., Morrison, D., Turco, R. P. and Covey, C., 1997, Environmental perturbations caused by the impacts of asteroids and comets. *Rev. Geophys.* **35**(1): 41–78.

Tozer, D. C., 1972, The present thermal state of the terrestrial planets. *Phys. Earth Planet. Inter.* **6**: 182–197.

Tromp, J., 1993, Support for anisotropy of the Earth's inner core from splitting in free oscillation data. *Nature* **366**: 678–681.

Turcotte, D. L. and Schubert, G., 2002, *Geodynamics*. Cambridge: Cambridge University Press.

Turner, G. M. and Thompson, R., 1981, Lake sediment record of the geomagnetic secular variation in Britain during Holocene times. *Geophys. J. R. Astron. Soc.* **65**: 703–725.

Utsu, T., 1961, A statistical study of the occurrence of aftershocks. *Geophys. Magazine* **30**: 521–605.

Utsu, T., 2002, Statistical features of seismicity. In *International Handbook of Earthquake Engineering and Seismology*, ed. W. H. K. Lee. San Diego: Academic Press. Part A, pp. 719–732.

Van der Voo, R, 1990, Phanerozoic poles from Europe and North America and comparisons with continental reconstruction. *Rev. Geophys.* **28**: 167–206.

Van der Voo, R., 1992, *Paleomagnetism of the Atlantic, Tethys and Iapetus Oceans*. Cambridge: Cambridge University Press.

Vanyo, J. D., 1991, A geodynamo powered by lunisolar precession. *Geophys. Astrophys. Fluid Dyn.* **59**: 209–234.

Vashchenko, V. Ya., and Zubarev, V. N., 1963, Concerning the Grüneisen constant. *Sov. Phys. Solid State* **5**: 653–655.

Veizer, J. and Jansen, S. L., 1979, Basement and sedimentary recycling and continental evolution. *J. Geol.* **87**: 341–370.

Veizer, J. and Jansen, S. L., 1985, Basement and sedimentary recycling – 2: Time dimension to global tectonics. *J. Geol.* **93**: 625–643.

Venkataraman, A. and Kanamori, H., 2004, Observational constraints on the fracture energy of subduction zone earthquakes. *J. Geophys. Res.* **109B**:5302. doi:10.1029/2003JB002549.

Vidale, J. E., 2001, Peeling back the layers in Earth's mantle. *Science* **294**: 313.

Vidale, J. E., Dodge, D. A. and Earle, P. S., 2000, Slow differential rotation of the Earth's inner core indicated by temporal changes in scattering, *Nature* **405**: 445–448.

Vidale, J. E. and Earle, P. S., 2000, Fine-scale heterogeneity in the Earth's inner core, *Nature* **405**: 273–275.

Vidale, J. E., and Hedlin, M. A. H., 2000, Evidence for partial melt at the core–mantle boundary north of Tonga from the strong scattering of seismic waves, *Nature* **391**: 682–685.

Vine, F. J. and Matthews, D. H., 1963, Magnetic anomalies over ocean ridges. *Nature* **199**: 947–949.

Vinet, P., Ferrante, J., Rose, J. H. and Smith, J. R., 1987, Compressibility of solids. *J. Geophys. Res.* **92**: 9319–9325.

Vondrák, J., 1999, Earth rotation parameters, 1899.7–1992.0, after reanalysis within the Hipparcos frame. *Surveys in Geophys*. **20**: 169–195.

Wasson, J. T., 1985, *Meteorites: Their Record of Early Solar System History*. New York: Freeman.

Watson, E. B. and Harrison, T. M., 2005, Zircon thermometer reveals minimum melting conditions on earliest Earth. *Science* **308**: 841–844.

Watt, P. J., Davies, G. F. and O'Connell, R. J., 1976, The elastic properties of composite materials. *Rev. Geophys. Space Phys.* **14**: 541–563.

Watts, A. B., 2001, *Isostasy and Flexure of the Lithosphere*. Cambridge: Cambridge University Press.

Weaver, H. A., Stern, S. A., Mutchler, M. J., Steffl, A. J., Buie, M. W., Merline, W. J., Spencer, J. R., Young, E. F. and Young, L. A., 2006, Discovery of two new satellites of Pluto. *Nature* **439**: 943–945.

Webb, D. J., 1982, Tides and the evolution of the Earth-Moon system. *Geophys. J. R. Astron. Soc.* **70**: 261–271.

Weertman, J. and Weertman, J. R., 1992, *Elementary Dislocation Theory*. Oxford: Oxford University Press.

Wetherill, G. W., 1968, Stone meteorites: time of fall and origin. *Science* **159**: 79–82.

Wetherill, G. W., 1981, Nature and origin of basin-forming projectiles. *Proc. Lunar Plan. Sci.* **12A**: 1–18.

Wetherill, G. W., 1985, Asteroidal source of ordinary chondrites. *Meteoritics* **20**: 1–22.

Whaler, K. A., 1980, Does the whole of the Earth's core convect? *Nature* **287**: 528–530.

Wheeler, K. T., Walker, D., Fei, Y., Minarik, W. G., and McDonough, W. F., 2006, Experimental partitioning of uranium between liquid iron sulfide and liquid silicate: implications for radioactivity in the Earth's core. *Geochim. Cosmochim. Acta* **70**: 1537–1547.

White, M. L., 1972, Jet streams and the development of the solar system. *Nature Phys. Sci.* **238**: 104–105.

Wielandt, E., 1987, On the validity of the ray approximation for interpreting delay times. In Nolet, G. (ed.), *Seismic Tomography*. Dordrecht: Reidel, pp. 85–98.

Wiens, D. A. and Stein, S., 1985, Implications of oceanic intraplate seismicity for plate stresses, driving forces and rheology. *Tectonophysics* **116**: 143–162.

Wignall, P. B., 2001, Large igneous provinces and mass extinctions. *Earth Sci. Rev.* **53**: 1–33.

Williams, G. E., 1990, Tidal rhythmites: key to the history of the Earth's rotation and the lunar orbit. *J. Phys. Earth* **38**: 475–491.

Williams, G. E., 1991, Milankovitch-band cyclicity in bedded halite deposits contemporaneous with Late Ordovician–Early Silurian glaciation, Canning Basin, Western Australia. *Earth Plan. Sci. Lett.* **103**: 143–155.

Williams, G. E., 2000, Geological constraints on the preCambrian history of the Earth's rotation and the Moon's orbit. *Rev. Geophys.* **38**: 37–59.

Williams, J. G., Ratcliffe, J. T. and Boggs, D. H., 2004, Lunar rotation orientation and science. *EOS (Trans. AGU Fall Meeting Supplement)* **85**(47) F603 (Abstract G33A-08).

Williams, Q., Revenaugh, J. and Garnero, E., 1998, A correlation between ultra-low basal velocities in the mantle and hot spots. *Science* **281**: 546–549.

Willson, R. C. and Hudson, H. S., 1991, The sun's luminosity over a complete solar cycle. *Nature* **351**: 42–44.

Wilson, R. L., 1962, The palaeomagnetism of baked contact rocks and reversals of the Earth's magnetic field. *Geophys. J. R. Astron. Soc.* **7**: 194–202.

Wilson, R. L., 1970, Permanent aspects of the Earth's non-dipole magnetic field over Upper Tertiary time. *Geophys. J. R. Astron. Soc.* **19**: 417–437.

Wisdom, J., 1983, Chaotic behaviour and the origin of the 3/1 Kirkwood gap. *Icarus* **56**: 51–74.

Wolbach, W. S., Lewis, R. S. and Anders, E., 1985, Cretaceous extinctions: evidence for wildfire and search for meteoritic material. *Science* **230**: 167–170.

Wood, B. J., 1993, Carbon in the core. *Earth Plan. Sci. Lett.* **117**: 593–607.

Wood, J. A., 1964, The cooling rates and parent planets of several iron meteorites. *Icarus* **3**: 429–459.

Woodhouse, J. H., 1983, The joint inversion of seismic waveforms for lateral variations in Earth structure and earthquake source parameters. In Kanamori, H. and Boschi, E. eds. *Proceedings of the Enrico Fermi International School of Physics*, **85**. Amsterdam: North Holland, pp. 366–397.

Woodhouse, J.H, 1988, The calculation of eigenfrequencies and eigenfunctions of the free oscillations of the earth and the sun. In Doornbos, D.J. (ed.), *Physics of the Earth's Interior. Seismological Algorithms*, London: Academic Press.

Woodhouse, J.H. and Dziewonski, A.M., 1984, Mapping the upper mantle: three dimensional modelling of Earth structure by inversion of seismic waveforms. *J. Geophys. Res.* **89**: 5953–5986.

Woodhouse, J.H., Giardini, D. and Li, X.-D., 1986, Evidence for inner-core anisotropy from splitting in free oscillation data. *Geophys. Res. Lett.* **13**: 1549–1552.

Xie, S. and Tackley, P.J., 2004, Evolution of helium and argon isotopes in a convecting mantle. *Phys. Earth Planet. Inter.* **146**: 417–439.

Xu, F., Vidale, J.E. and Earle, P.S., 2003, Survey of precursors to P′P′: fine structure of mantle discontinuities. *J. Geophys. Res.* **108**(B1): 2024. doi:10.1029/2001JB000817,2003.

Yeganeh-Haeri, A., 1994, Synthesis and reinvestigation of the elastic properties of magnesium silicate perovskite. *Phys. Earth Planet. Inter.* **87**: 111–121.

Yin, A., 2000, Mode of east-west extension in Tibet suggesting a common origin of rifts in Asia during the Indo-Asian collision. *J. Geophys. Res.* **105**(B9): 21745–21759.

Yoder, C.F., Konopliv, A.S., Yuan, D.N., Standish, E M. and Folkner, W.M., 2003, Fluid core size of Mars from detection of the solar tide. *Science* **300**: 299–303.

Yoshida, M., 2004, Possible effects of lateral viscosity variations induced by plate tectonic mechanism on geoid inferred from numerical models of mantle convection. *Phys. Earth Planet. Inter.* **147**: 67–85.

Yoshida, S., Sumita, I. and Kumazawa, M., 1996, Growth model of the inner core coupled with outer core dynamics and the resulting elastic anisotropy. *J. Geophys. Res.* **101**: 28085–28103.

Young, C.J., and Lay, T., 1990, Multiple phase analysis of the shear velocity structure in the D″ region beneath Alaska. *J. Geophys., Res.* **95**: 17385–17402.

Yuan, X. *et al.* (22 authors), 2000, Subduction and collision processes in the central Andes constrained by converted seismic phases. *Nature* **408**: 958–961.

Yukutake, T., 1989, Geomagnetic secular variation: theory. In James (1989), pp. 578–584.

Yukutake, T. and Tachinaka, T., 1969, Separation of the Earth's magnetic field into the drifting and the standing parts. *Bull. Earthquake Res. Inst. Univ. Tokyo* **47**: 65–97.

Zeng, Y., 1993, Theory of scattered P- and S-wave energy in a random isotropic scattering medium. *Bull. Seism. Soc. Am.* **83**(4): 1264–1276.

Zhang, J., Song, X., Li, Y., Richards, P.G., Sun, X. and Waldhauser, F., 2005, Inner core differential motion confirmed by earthquake waveform doublets. *Science* **309**: 1357–1360.

Zhang, Q., Soon, W.H., Baliunas, S.L., Lockwood, G.W., Skiff, B.A. and Radick, R.R., 1994, A method of determining possible brightness variations of the Sun in past centuries from observations of solar-type stars. *Astrophys. J.* **427**: L111–L114.

Zhang, Y.S. and Lay, T., 1999, Evolution of oceanic upper mantle structure. *Phys. Earth Planet. Inter.* **114**: 71–80.

Zhang, Y.S. and Tanimoto, T., 1991, Global Love wave phase velocity variation and it significance to plate tectonics. *Phys. Earth Planet. Inter.* **66**: 160–202.

Zhang, Y.S. and Tanimoto, T., 1992, Ridges, hotspots and their interaction as observed in seismic velocity maps. *Nature* **335**: 4–49.

Zho, W. *et al.* (15 authors), 2001, Crustal structure of central Tibet as derived from Project INDEPTH wide-angle seismic data. *Geophys. J. Int.* **145**: 486–498.