Astronomy and Astrophysics Library

Eugene F. Milone
William J.F. Wilson

Solar System Astrophysics

Background Science and the Inner Solar System

Second Edition

Springer

Astronomy and Astrophysics Library

For further volumes:
http://www.springer.com/series/848

Eugene F. Milone • William J.F. Wilson

Solar System Astrophysics

Background Science and the Inner Solar System

Second Edition

Springer

Eugene F. Milone
Professor Emeritus
Dept. Physics & Astronomy
University of Calgary
Calgary, Alberta
Canada

William J.F. Wilson
Senior Instructor Emeritus
Dept. Physics & Astronomy
University of Calgary
Calgary, Alberta
Canada

Cover Illustration: “Mysterium Cosmographicum” by David R. Mouritsen

ISSN 0941-7834 ISSN 2196-9698 (electronic)
ISBN 978-1-4614-8847-7 ISBN 978-1-4614-8848-4 (eBook)
DOI 10.1007/978-1-4614-8848-4
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2013949766

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Preface to the First Edition

This work is appearing in two parts because its mass is the result of combining detailed exposition and recent scholarship. Book I, dealing mainly with the inner solar system, and Book II, mainly on the outer solar system, represent the combined, annually updated, course notes of E. F. Milone and W. J. F. Wilson for the undergraduate course in solar system astrophysics that has been taught as part of the Astrophysics Program at the University of Calgary since the 1970s. The course, and so the book, assumes an initial course in astronomy and first-year courses in mathematics and physics. The relevant concepts of mathematics, geology, and chemistry that are required for the course are introduced within the text itself.

Solar System Astrophysics is intended for use by second- and third-year astrophysics majors, but other science students have also found the course notes rewarding. We therefore expect that students and instructors from other disciplines will also find the text a useful treatment. Finally, we think the work will be a suitable resource for amateurs with some background in science or mathematics. Most of the mathematical formulae presented in the text are derived in logical sequences. This makes for large numbers of equations, but it also makes for relatively clear derivations. The derivations are found mainly in Chaps. 2–6 in the first volume, *Background Science and the Inner Solar System,* and in Chaps. 10 and 11 in the second volume, *Planetary Atmospheres and the Outer Solar System.* Equations are found in the other chapters as well but these contain more expository material and recent scholarship than some of the earlier chapters. Thus, Chaps. 8 and 9, and 12–16 contain some useful derivations, but also much imagery and results of modern studies.

The first volume starts with a description of historical perceptions of the solar system and universe, in narrowing perspective over the centuries, reflecting the history (until the present century, when extra-solar planets again have begun to broaden our focus). The second chapter treats the basic concepts in the geometry of the circle and of the sphere, reviewing and extending material from introductory astronomy courses, such as spherical coordinate transformations. The third chapter then reviews basic mechanics and two-body systems, orbital description, and the computations of ephemerides, then progresses to the restricted three-body and n-body cases, and concludes with a discussion of perturbations. The fourth chapter treats the core of the solar system, the Sun, and is not a bad introduction to solar or stellar astrophysics; the place of the Sun in the galaxy and in the context of other stars is described, and radiative transport, optical depth, and limb-darkening are introduced. In Chap. 5, the structure and composition of the Earth are discussed, the Adams–Williamson equation is derived, and its use for determining the march of pressure and density with radius described. In Chap. 6, the thermal structure and energy transport through the Earth are treated, and in this chapter the basic ideas of thermodynamics are put to use. Extending the discussion of the Earth's interior, Chap. 7 describes the rocks and minerals in the Earth and their crystalline structure. Chapter 8 treats the Moon, its structures, and its origins, making use of the developments of the preceding chapters. In Chap. 9, the surfaces of the other terrestrial planets are described, beginning with Mercury. In each of the three sections of this chapter, a brief historical discussion is followed by descriptions of modern ground-based and space mission results, with some of the spectacular imagery of Venus and Mars. The chapter concludes with a description of the evidence for water and surface modification on Mars. This concludes the discussion of the inner solar system.

The second volume begins in Chap. 10 with an extensive treatment of the physics and chemistry of the atmosphere and ionosphere of the Earth and an introduction to meteorology, and this discussion is extended to the atmospheres of Venus and Mars. Chapter 11 treats the magnetospheres of these planets, after a brief exposition of electromagnetic theory. In Chap. 12, we begin to treat the outer solar system, beginning with the gas giants. The structure, composition, and particle environments around these planets are discussed, and this is continued in Chap. 13, where the natural satellites and rings of these objects are treated in detail, with abundant use made of the missions to the outer planets. In Chap. 14, we discuss comets, beginning with a historical introduction that highlights the importance of comet studies to the development of modern astronomy. It summarizes the ground- and space-based imagery and discoveries, but makes use of earlier derivations to discuss cometary orbits. This chapter ends with the demise of comets and the physics of meteors. Chapter 15 treats the study of meteorites and the remaining small bodies of the solar system, the asteroids *(aka* minor planets, planetoids), and the outer solar system "Kuiper Belt" objects, and the closely related objects known as centaurs, plutinos, cubewanos, and others, all of which are numbered as asteroids. The chapter ends with discussions of the origin of the solar system and of debris disks around other stars, which point to widespread evidence of the birth of other planetary systems. Finally, in Chap. 16, we discuss the methods and results of extra-solar planet searches, the distinctions among stars, brown dwarfs, and planets, and we explore the origins of planetary systems in this wider context.

At the end of nearly every chapter we have a series of challenges. Instructors may use these as homework assignments, each due 2 weeks after the material from that chapter were discussed in class; *we* did! The general reader may find them helpful as focusing aids.

Preface to the Second Edition

As in the first edition, we maintain the two-volume bifurcation of the inner and outer regions of the solar system. In the first volume, we again begin with a historical overview but expand the horizon to include glimpses of extra-solar planetary systems. The basic mathematics, mechanics, geophysics, thermodynamics, chemistry, astrophysics, and mineralogical principles required for a sound introduction to space science have been revised with improved illustrations and examples drawn from wider sources. In Chap. 4, we have added descriptions of the features of the active Sun. Chapter 8, on the Moon, has been updated with results of probes of water at the poles and a fresh discussion of the Moon's origin. In Chap. 9, the Messenger mission has provided vital new details about Mercury, and the history of the study of Venus has been expanded. The Mars section includes results from the Curiosity mission and a description of current views of the search for life in the Viking mission. The crustal changes in Mars since its formation, and an enlarged discussion of climate changes, expand that section further. Similar expansions of the chapters of the second volume have vastly expanded the discussions of atmospheres, magnetospheres, the gas and ice giants of the outer solar system and their moons and ring systems. The discussion of meteors and meteorite impacts has been enlivened by recent events, and a deepening understanding of the role played by disks in the early history of planetary formation. The burgeoning field of extrasolar planets has been reflected in the vastly increased discussion in the last chapter, with the increasing knowledge of the properties of extrasolar planets and their more massive siblings, the brown dwarfs. The dynamical interactions being studied with increasingly sophisticated software simulations have greatly illuminated the likely dynamical development of the solar system. As in all such investigations, present questions have been and are being answered, but new puzzles arise, and it is the anticipation of the new adventures required to explain them that makes this field truly exciting.

Calgary, Alberta, Canada

Eugene F. Milone
William J.F. Wilson

Acknowledgments

These volumes owe their origin to more than 30 years of solar system classes in the Astrophysics Program at the University of Calgary, called, at various times, Geophysics 375, Astrophysics 301, 309, and 409. Therefore, we acknowledge, first, the students who took these courses and provided feedback. It is also a pleasure to thank the following people for their contributions:

David R. Mouritsen, formerly of Calgary and now in Toronto, provided for Chap. 1 and our covers an image of his original work of art, an interpretation of Kepler's *Mysterium Cosmographicum,* in which the orbits of the planets are inscribed within solid geometric figures.

In Chap. 3, the Bradstreet and Steelman software package *Binary Maker 3* was used to create an image to illustrate restricted three-body solutions.

University of Calgary Professor Emeritus Alan Clark gave us an image of an active region and detailed comments on the material of Chap. 4 for both editions; Dr. Rouppe van der Voort of the University of Oslo provided high-quality images of two other active region figures, for Chap. 4; the late Dr. Richard Tousey of the US Naval Research Laboratory provided slides of some of the images, subsequently scanned for Chap. 4; limb-darkened spectral distribution plots were provided by Dr. Robert L. Kurucz, of the Harvard-Smithsonian Center for Astrophysics; Dr. Charles Wolff, of Goddard Space Flight Center, NASA, reviewed the solar oscillations sections and provided helpful suggestions for both editions; Dr. David H. Hathaway, Marshall Space Flight Center, NASA, provided permission to use his solar cycle sunspot area graphic.

Dr. David Eaton of the Geoscience Department of the University of Calgary made available for Chap. 5 his seismogram image of the major Japanese earthquake of 2011.

Mr. Gerhard Brunthaler provided inspiration for the derivation of the contribution of tidal heating to the Earth's interior heat budget in Chap. 6.

Dr. D. J. Stevenson provided helpful criticism of our lunar origins figures, and Dr. Robin Canup kindly prepared panels of her lunar simulations for our Fig. 8.13.

Dr. James A. Slavin, of the Dept. Atmospheric, Oceanic and Space Sciences, University of Michigan, critiqued our discussion of Mercury's magnetosphere in Chap. 9. We also acknowledge with thanks and gratitude the help of Prof. S. van Gasselt, Inst. Geologische Wissenschaften, Freie Univ., Berlin and Ms. Stefanie Potts, Sekretariat Planetelogie, for their kindness in providing and enabling our use of Fig. 9.51, and Ms. Nadia Imbert-Vier of ESA for her helpful and timely responses for permission to use ESA images. Dr. Andrew Yau provided excellent notes as a guest lecturer in Asph 409 on the Martian atmosphere and its evolution, which contributed to our knowledge of the material presented in Chaps. 9, 10, and 11; similarly, lectures by Professor J. S. Murphree of the University of Calgary illuminated the magnetospheric material described in Chap. 11. For the second edition, Professor David Knudsen of the University of Calgary critiqued our expanded Chap. 11.

NASA's online photo gallery provided many of the images in Chaps. 8, 9, 12, 13, 14, and some of those in Chap. 15; additional images were provided by the Naval Research Laboratory (of both the Sun and the Moon). Some of these and other images involved work by other institutions, such as the U.S. Geological Survey, the Jet Propulsion Laboratory, Arizona State Univ., Cornell, the European Space Agency, the Italian Space Agency (ASI), CICLOPS, CalTech, Univ. of Arizona, Space Science Institute, Boulder, the German Air and Space Center (DLR), Brown University, the Voyagers and the Cassini Imaging Teams, the Hubble Space Telescope, University of Maryland and the Deep Impact mission team, the Minor Planet Center, Applied Physics Laboratory of the Johns Hopkins University, and the many individual sources, whether cited in captions or not, who contributed their talents to producing these images.

Dr. John Trauger provided a high-resolution UV image of Saturn and its auroras for Chap. 12, and Dr. Chris Arridige (Univ. College London) provided helpful suggestions for our magnetospheric effects in the Saturn system in Chap. 12. Dr. Stuart Eves provided an important reference to William Herschel's reports of his observations for Chap. 13.

Dr. William Reach, Caltech, provided an infrared mosaic image of Comet Schwassmann–Wachmann 3, and Mr. John Mirtle of Calgary provided many of the comet images for Chap. 14, including those of Comets 109P/Swift-Tuttle, C/1995 O1 (Hale–Bopp), C/Hyakutake, Lee, C/Ikeya–Zhang, Brorsen-Metcalfe, and Machholz; Professor Michael F. A'Hearn of the University of Maryland for his critique of the comet content of Chap. 14. Univ, of Calgary Prof. Emer. Alan Clark and Dr. William Livingstone, KPNO, Tucson, provided images of meteors. Mr. Mike Hankey of the American Meteor Society provided useful sources of recent fireball frequency data.

Dr. Allan Treiman, of the Lunar and Planetary Institute, Houston, was kind enough to provide background material on the debate over the ALH84001 organic life question, mentioned in Chap. 15. Mr. Matthias Busch made available his *Easy-Sky* images of asteroid family distributions, and the Minor Planet Center provided the high-resolution figures of the distributions of the minor planets in the inner and outer solar system.

Dr. Charles Lineweaver, University of New South Wales, provided a convincing illustration for the brown dwarf desert, illustrated in Chap. 16; University of Calgary graduate student Michael Williams provided several figures from his MSc thesis on the extra-solar planet HD 209458b, the first to be found transiting a star, for Chap. 16.

Mr. Alexander Jack assisted in updating and improving the readability of equations and text in some of the early chapters, and he and Ms. Veronica Jack assisted in developing the tables of the extra-solar planets and their stars for Chap. 16 in both editions.

In addition, we thank the many authors, journals, and publishers who have given us permission to use their figures and tabular material or adaptations thereof, freely. Finally, it is also a pleasure to thank Springer editors Dr. Hans Koelsch, Dr. Harry Blom, and their associates, Christopher Coughlin, Brian Halm, and Jennifer Satten, for their support for this project.

Contents

Chapter 1
Perceptions of the Solar System in History

1.1 Evolving Perceptions

The solar system has been around for a long time! Our perceptions of it, on the other hand, date back, arguably, only to the Upper Paleolithic (~70,000 to ~10,000 years ago).

The Paleolithic evidence for interest in the Moon, for example, is in the form of possible tallies of days in such features as the 17-in. high sculpture called the *Venus of Laussel* in a rock shelter dated from 20,000 to 18,000 y BP in the Dordogne region of France (see Campbell 1988, pp. 65–66 for a discussion of its symbolic significance and Marshack 1972, p. 335 for arguments for its use as a tally) and the Blanchard bone (among other artifacts) with a complicated chain of crescent incisions, also investigated by Marshack. The amply endowed *Venus* holds in her right hand an upturned horn on which are incised 13 grooves (defining 14 non-groove areas), a number said to represent an approximation to the number of waxing crescents in a year (12⅓) and the number of days between new and full moon (~14¾) (Marshack 1972).

In the Neolithic (or "New Stone Age," roughly from 6500 to 1500 BC), the evidence for the importance of the Sun and the Moon, at least, is overwhelming. It can be found in the many ancient alignment sites in the British Isles and is echoed in Stone Age cultures around the world, at least according to some interpretations. See Thom (1972) for a flavor of that evidence or, for example, Kelley and Milone (2011, Chap. 6) for a more recent summary.

Aside from practical astronomy, with calendrical usefulness for both agriculture and religions, astrology plays an increasing role in late antiquity (first several centuries AD). For the past four millennia, and especially during Hellenic and Hellenistic times (prior to and after Alexander the Great, respectively), we do know what people thought regarding the nature and origin of the solar system. Without attempting detailed examination of each one, we can characterize the principal theories about the solar system (in earlier times, the entire cosmos) as shown in Table 1.1.

E.F. Milone and W.J.F. Wilson, *Solar System Astrophysics: Background Science and the Inner Solar System*, Astronomy and Astrophysics Library,
DOI 10.1007/978-1-4614-8848-4_1, © Springer Science+Business Media New York 2014

Table 1.1 Ancient theories of the solar system

Pythagoras (~6th BC): Earth at centre of planetary spheres (included Sun, Moon and the sphere of the fixed stars)

Anaxagoras of Clazomenae (c. 500–428 BC): The Sun is a hot iron mass (based on meteorite evidence), bigger than the Peloponnesus; the Moon is a stone; the Earth, the centre of a cosmic vortex

Philolaus (~5th c. BC): The Earth moves, while the stellar sphere is immobile. He felt that there should be ten planets (fixed star sphere included), so Philolaus invented an anti-Earth, perpetually located between the Earth and a central fire about which all the planets, including the sun, moved; orbits were circular, but not coplanar. This is the earliest recorded theory to consider the Earth as a moving object—but it was to explain the daily western movement of the "fixed" stars and other objects, not the annual motions of the Sun or the planets

Eudoxus of Cnidus (~408–355 BC): The fixed stars and each planet are carried on separate, concentric, rotating spheres, on various axes, centered on the Earth

Aristarchus of Samos (~250 BC): The Sun is at the center; the Earth both revolves around the Sun and rotates on its own axis; the Moon revolves around the Earth

Apollonius of Perga (~220 BC): Combinations of motion in circular orbits

Hipparchus (2nd c. BC): The Earth is at the center; planets (including the Sun) revolve around the Earth; orbits are circular but non-concentric

Claudius Ptolemy (2nd c. AD): The Earth is at the center; planets (including the Sun) revolve around the Earth; the orbits are combinations of circular motions, characterized by deferent orbits and epicyclic gyrations

Origen (3rd c. AD): There exist a multiplicity of worlds, with the creation, fall, and redemption occurring on each

Martianus Capella (≲ 5th *c. ?)*: His *Satyricon* refers to a Sun-centered solar system; this popular work kept the notion alive in the West to Copernicus' time

Aryabhata (b. 476 AD): He allowed the possibility of a heliocentric universe (but his work was not known in the West until after Copernicus)

Most of the many notable figures in Tables 1.1 and 1.2 are discussed in Kelley and Milone (2011). Here, we single out only three for further discussion.

Martianus Capella was a poet and summarizer, who wrote his allegorical poem *Satyricon* after the sack of Rome by the Huns in 410 AD, which he mentions, and possibly before 429 AD, when Carthage was overrun by Vandals, which he does not. His description of a quasi-heliocentric system (like Tycho's model, it had Mercury and Venus orbiting the Sun), kept alive this idea. Copernicus explicitly mentions Capella's (and not Aristarchus') discussion of the heliocentric system.

Much later, following the Renaissance and Reformation, Tycho Brahe and Johannes Kepler were important transition figures.

Brahe himself contributed to a break in the classical paradigm by demonstrating, with observational data, that comets moved among the orbits of planets, thus shattering once and for all the notion that rotating crystalline spheres bore the planets. His discovery of a supernova and his determination that it was a very distant object demolished the idea of the immutability of the heavens. Moreover, this and his cometary discoveries refuted the ideas of Aristotle, for centuries considered the highest authority on scientific questions.

Kepler's early notion of the heliocentric planetary orbits carried on (crystalline) spheres inscribing and inscribed by the five regular polyhedral solids (see Fig. 1.1)

Table 1.2 Post-medieval, pre-nineteenth century theories of the solar system

Nicholas of Cusa (1401–1464): He is said to have championed a Sun-centered theory; no explicit writings

Nicholas Copernicus (1473–1543): Sun-centered solar system; planets moved (as classically) in circular orbits

Tycho Brahe (1546–1601): Sun-centred planetary scheme—but Sun and planets revolve about the Earth

Johannes Kepler (1571–1630): Elliptical orbits; this is the first explicit departure from circular orbits. Keplerian empirical "laws"

Rene Descartes (1596–1650): The solar system is a complex of vortices; moons and planets arise from vortices within vortices

Isaac Newton (1642–1727): Planetary orbital motion due to gravity. The solar system is far from the stars (considered distant because of lack of parallax and relative motions) which are themselves, therefore, suns

Georges-Louis Leclerc Buffon (1707–1788): Collisional origin for the solar system (Sun with comet)

Immanuel Kant (1724–1794) and Simon de Laplace (1749–1827): The solar system had a nebular origin; contraction and conservation of angular momentum caused disk formation

Ernst Florenz Friedrich Chladni (1756–1827): The early aggregation of dust became planetesimals, and some of these, planets (Chladni 1794)

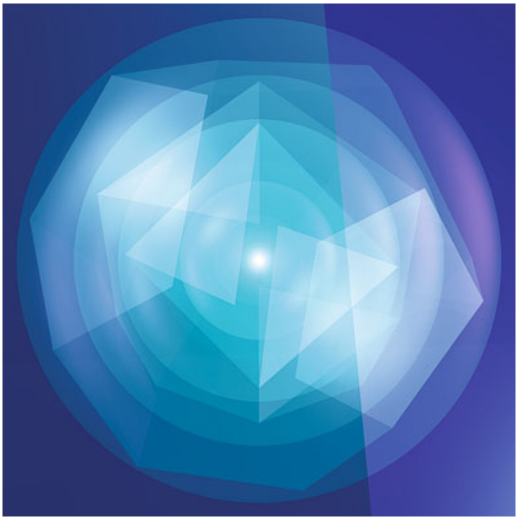


Fig. 1.1 An early Keplerian view of the solar system, inspired by Kepler's *Mysterium Cosmographicum* model of nested spheres and geometric solids. Original art by David R. Mouritsen (2005) and reproduced here with permission

as expressed in the first half of his *Mysterium Cosmographicum* (Kepler 1596), evolved over his lifetime into a realization that the orbits were ellipses produced by forces that depended on the distance from the Sun. His persistence in trying to make sense of Tycho Brahe's highly precise data led to his conclusion that planetary orbits could not be circular. The consequences of this profound discovery resulted in the "Breaking of the Circle," in many ways (Nicholson 1950).

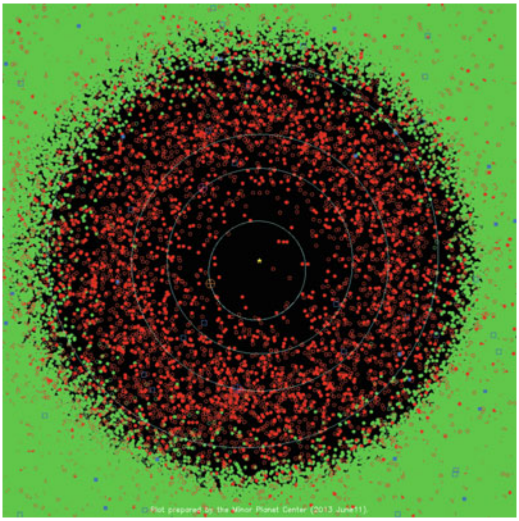


Fig. 1.2 The innermost part of the solar system showing the orbits of Mercury, Venus, Earth and Mars, along with the positions of those planets (*large circles* with *crosses*), minor planets (*red* and *green circles*), and comets (*blue squares*) on June 11, 2013. Well-observed small bodies are shown with *filled symbols*, others with *open symbols*. The *red circles* indicate minor planets that approach the Sun to within 1.3 au. The planets move CCW over time. Courtesy Gareth Williams (Minor Planet Center, SAO, Cambridge, MA)

Thus their pursuit of the highest quality observational data and unflinching belief in the meaningfulness of those data led both of them to renounce the geocentric universe, although Brahe's was a last effort to incorporate the idea of a stationary Earth into an empirically defensible model.

In the eighteenth and early nineteenth centuries, two principal ideas began to emerge: disk formation in a spinning nebula (Kant and LaPlace) and accretion through collisions with lesser bodies (Chladni).

The existence of lesser bodies in the solar system has been demonstrated over and over again, as the Earth is continually approached by comets, minor planets (also known as asteroids), and lesser objects down to dust grain sizes. Figure 1.2 displays the innermost part of the solar system as it appeared on June 11, 2013, as plotted by the International Astronomical Union's Minor Planet Center in Cambridge, Massachusetts. The orbits of Mercury, Venus, Earth, and Mars are shown along with comets (blue squares) and asteroids (circles). The red circles are asteroids that come within 1.3 au of the Sun. Well observed small bodies are represented by filled symbols, others by open symbols, the planets by large open circles with crosses, and the Sun by a gold star. The direction of motion of the planets and most of the other depicted objects is counterclockwise.

There are many nineteenth and twentieth century theories. Most of these theories involve either collisions or accretions or both. Table 1.3 presents some examples.

Table 1.3 Nineteenth and twentieth century theories of the origin of the solar system

A. W. Bickerton (1842–1929): Star-Sun collision; explosive eruption forms planets

R. A. Proctor (1837–1888): Planetesimal aggregations [Proctor (1898)]

T. C. Chamberlain (1843–1928): Star-Sun collision; tidal eruption creates planets; [Chamberlain (1904)]: Planetesimals

F. R. Moulton (1872–1952): Star-Sun collision; tidal eruption + planetary accretion [Moulton (1905)]

K. O. B. Birkeland (1867–1912): Ions in solar atmosphere form rings in solar magnetic field

S. A. Arrhenius (1859–1927): Direct Sun-star collision, leaving the Sun and long filament as remnants

H. Jeffreys (1891–1989): Grazing Sun-star collision, leaving long filament that fragmented

J. H. Jeans (1877–1946): Star-Sun collision producing tidal filament

H. P. Berlage (1856–1934): Solar particle emission lead to gaseous rings/disks

H. N. Russell (1877–1957): Binary star component disrupted, forming a filament [Russell (1935)]

D. ter Haar (1919–2002): Contracting, turbulent solar envelope developed into planets

H. O. G. Alfvén (1908–1995): Sun collided with a gas cloud which became ionized, and formed rings in the Sun's magnetic field; electromagnetic braking and transfer of angular momentum

O. I. Schmidt (1891–1956): Sun collided with a swarm of interstellar bodies which became planets by accretion; refined by R. A. Lyttleton (1911–1995)

C. F. von Weizsäcker (1912–2007): Turbulent eddies in protosun formed planets and satellites [von Weizsäcker (1943)]

F. Hoyle (1915–2001): Sun's binary companion went supernova, producing gaseous shells; remnant star left the system

F. Whipple (1906–2004): Protosun captured dust cloud of large angular momentum

G. Kuiper (1905–1973): Gravitational instabilities in protosun's gaseous envelope became planets [Kuiper (1951)]

V. S. Safronov (1917–1999): Aggregation of dust into planetesimals [Safronov (1969)]

A. G. W. Cameron (1924–2005): Gaseous protoplanet theory [Cameron (1978)]

C. Hayashi (1920–2010): Aggregation into planetesimals ("Kyoto" school) [Hayashi et al. (1985)]

Several of the theories, including the most recent, are cited in the references list. Note the trend from Sun-star collisional theories to accretion theories in this interval.

Any thorough study of the solar system draws from chemistry, geology, and even biology, and numerous insights from those sciences will be brought into and used in this book. But, it is still basically astronomy. Observational astronomy has provided the basic data which are needed to understand the planets and other objects of the solar system, even if most of the new critical data now come from satellites and space probes, containing not only imaging cameras, but spectrographs (exploring the spectral energy distribution from radio and infrared to x-rays), magnetometers (to probe the structure and strength of magnetic fields), and particle detectors. But the latter does emphasize that remote sensing plays a vital role, and therefore our understanding of the solar system is more and more through "space science."

When we consider that the ultimate quest is to understand how the solar system came into being, how it evolves, and to what end, it is clear that a critical field of investigation has to be solar system dynamics. Therefore, this is an important area of study even for those who are not going to work for the Canadian, US, and European space agencies (CSA, NASA, and ESA, respectively) or for any of the space agencies developed in other countries around the world. We will take up this subject in a later chapter.

1.2 Striking Facts About the Solar System

The investigation of the nature of the solar system points to several striking facts:

- The solar rotation and the revolution of all the planets are in the same sense: CCW as viewed from the north ecliptic pole (NEP).
- The orbits are very nearly coplanar (the biggest departures being for the innermost and the outermost planetary objects—Mercury and Pluto); and, again except for Mercury and dwarf and minor planets, very nearly circular.
- The spacing of the planets is not random, but is described to a degree by the Titius-Bode "law" (Table 1.4) and similar laws. Some aspects of these "laws" are discussed below.
- Although the mass is strongly concentrated in the Sun, the angular momentum is not.
- The coplanar revolutions of the planets and the solar rotation (to be discussed in Chap. 4) already make a disk formation of the solar system more likely than a collisional origin.
- The low orbital eccentricities of the planets strengthen the case. The circularity of Neptune's orbit, the outermost and thus least strongly bound of all the major planets (Pluto, Eris, and other "dwarf planets" excepted from this category), is especially compelling.

Let us begin our narrative by examining one of the most mysterious of the striking facts: the quasi-predictable spacings of the planets. The Titius-Bode law (Bode 1772; Wurm 1787; Jaki 1972; Nieto 1972) can be expressed in the form:

$$r = (3 \times 2^n + 4)/10 \qquad (1.1)$$

where r is the average or mean distance of the planet from the Sun in units of the Earth's mean distance, the astronomical unit (au), and $n = -\infty$, 0, 1, 2, 3, ... 6 (7–9, Neptune-Eris, are not well represented). The lack of a clearcut physical

Table 1.4 The Titius-Bode "law"

	$r = 2^n \cdot 0.3 + 0.4$		
Planet	n	Prediction	True r
Mercury	$-\infty$	0.4	0.39
Venus	0	0.7	0.72
Earth	1	1.0	1.00
Mars	2	1.6	1.52
Minor planets	3	2.8	<2.8>
Jupiter	4	5.2	5.20
Saturn	5	10.0	9.54
Uranus	6	19.6	19.18
Neptune	7	38.8	30.07
Pluto	8	77.2	39.46
Eris	9	154.0	67.78

Table 1.5 Blagg-Richardson Fitting Parameters for Select Planetary Systems

log r_n = log r_0 + n log a

System	N	log (r_0)	log a	r_0	a
Solar System	11	−0.428(60)	+0.2303(57)	+0.373(+55/−48)	+1.700(+23/−22)
HD 10180	9	−1.58(12)	+0.245(16)	+0.026(+9/−7)	+1.756(+66/−64)
Kepler 11	6	−1.082(56)	+0.135(13)	+0.083(+11/−10)	+1.365(+43/−42)
Kepler 33	5	−1.207(60)	+0.140(19)	+0.094(+14/−12)	+1.381(+61/−59)
Kepler 20	5	−1.399(89)	+0.210(28)	+0.040(+9/−7)	+1.62(+11/−10)
GJ 876	4	−1.43(23)	+0.38(10)	+0.038(+26/−15)	+2.42(+64/−50)

basis for the Titius-Bode relation and its failure to predict a correct mean distance for (at least) the outermost major planet, Neptune, indicate the term "law" is inappropriate. Therefore, although we use it and acknowledge the historical importance of the Titius-Bode "law," we insert quotation marks around "law" from this point forward.

We can seek more effective formulations also. The relation can be expressed, for example, in this newer form,

$$r_n = r_0 a^n \tag{1.2}$$

where r_n is the average distance in au of the nth planet (Mercury is $n = 0$), in order of distance, from the Sun, and where $a \equiv 1.73$ in the Blagg-Richardson formulation (see Nieto 1972). Note that a in (1.2) is a fitting parameter and not the semimajor axis, one of the orbit elements to be discussed in later chapters, and that one can either determine or assume the quantity r_0. Nieto (1972) successfully applies the relation to the principal satellites of the major planets of the solar system as well.

One may write (1.2) in logarithmic form as:

$$\log r_n = \log\ r_0 + n\ \log\ a \tag{1.3}$$

In Table 1.5 and Fig. 1.3 we present Blagg-Richardson relations in the logarithmic form (1.3) for six observed planetary systems, including the solar system. In this exercise, both of the parameters **log r_0** and **log a** are computed by least squares fitting to the observed data for each system using the regression algorithm of a spreadsheet program. In principle, this should allow a better fitting to be achieved than by assuming the base of the Blagg-Richardson logarithm, a = 1.73 (or in base 10 logarithms, assuming log a = 0.238). The quantities ***r*** and ***a*** in Table 1.5 and the calculated values of **log r_n** in Fig. 1.3 are then computed from the derived logarithmic parameters. **N** in Table 1.5 is the assumed number of planets for this calculation; in the case of the solar system, we have included dwarf planets. The uncertainties in the derived fitting parameters in Table 1.5 are given in parentheses in units of the last decimal place.

An examination of Fig. 1.3 suggests that such spacing or scaling laws seem to apply in other planetary systems also, if by "apply" we mean that the residuals are within ~5 %, although their Blagg-Richardson zero points and coefficients differ.

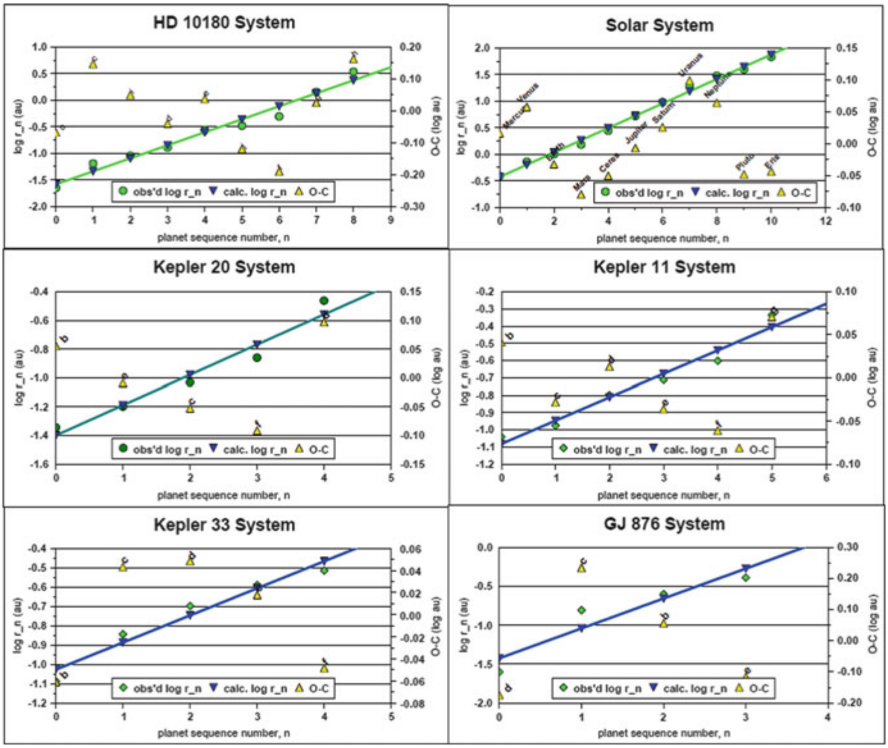


Fig. 1.3 Plots of the Blagg-Richardson relations in Table 1.5 for the solar system and selected extrasolar planet systems, compared to the observed values of log r_n

It is interesting that the coefficient for the HD 10180 planetary system, with nine known planets, agrees with that for the solar system. Scaling laws seem to work for the satellite systems of our major planets (more or less—Saturn's moon Mimas seems to be hard to fit—see Nieto (1972)), so perhaps it is not surprising that they should work for exoplanets. That they apply at all may tell us something about the distribution of protoplanetary disk densities and dynamical interactions among protoplanets and disks, and subsequently among planets. There is some evidence, for example, that the inner region of our solar system is dynamically "full" in the sense that when a test planet is introduced in the simulations, it is very likely to escape over much shorter timescales than the age of the solar system. The outer region is similarly said to be dynamically "full" and stable (reported by Perryman 2011). Perhaps applicability of Titius-Bode-type laws is in some way related to this "fullness." At present, our knowledge of disk structure and evolution, although growing rapidly, is still incomplete (see, e.g., Williams and Cieza 2011), and, of course, so is our knowledge of extrasolar planet systems, and our own system's true dynamical history. In the meantime, the current prevailing view is that Titius-Bode type "laws" illustrate primarily coincidence. Until our understanding of the issues we will be discussing throughout Solar System Astrophysics is more complete, and a rigorous and adequately robust theory is able to provide a basis for the relations, that view is certain to continue to prevail.

1.3 The Methods and Goals of Solar System Astrophysics

Our list of "striking facts" and other properties of the solar system will be reviewed at the beginning of Chap. 4 and again later, mainly in Chaps. 15 and 16 of Milone and Wilson (2014), when we consider the origins of our and other planetary systems. At the moment, we highlight the disk-like clues among those facts.

In our final chapter, we investigate the properties of extrasolar planets. The existence of disks around other stars and giant molecular clouds with which protostars are associated are further evidence for a disk origin of the solar system. Currently, planets are believed to arise from protostellar disks, but there are sharp disagreements over the separate roles of disk condensation through "gravitational instability" and accretion of other condensates or larger—perhaps pre-existing clumps of matter ("core-accretion"). The importance of disks is empirically based not only on infrared observations of tori seen around other stars (most famously, but far from exclusively, β Pictoris), but also on the basis of meteorite and theoretical studies. However, theoretical difficulties in explaining the formation of the lesser giants at their current locations in the solar system and the presence of "hot Jupiters" in other star systems strongly suggest dynamical migrations of planets from their points of origin, if the disk origin is to be sustained. Indeed, modern simulations provide mounting evidence that the lesser giants in the solar system, Uranus and Neptune, were formed closer to the Sun and were driven further out by dynamical interactions. In the last chapter of Milone and Wilson (2014) we will summarize what can be generalized about the origins of planetary systems and the prospects of finding life-sustaining terrestrial planets in other star systems.

As this chapter demonstrates, any study of the origin of the solar system must be a kind of mystery-solving expedition. So, we need to take note of the clues as we go along, as a police inspector in attempting to unscramble a forensic puzzle.

We start by providing some basic investigatory tools and then begin the search for clues in the dynamical and physical structure of the solar system and the increasing number of known extrasolar planetary systems.

Challenges

[1.1] Try to categorize the theories of the origin of the solar system, listing the theories below each category. Is there any evidence for historical evolution or evidence of progress among the theories of a particular type?

[1.2] Compare the computed distances of the Titius-Bode and Blagg-Richardson laws to the mean distances of the planets from the Sun. A spreadsheet is the most convenient way of doing this. Can you formulate another relation that describes these distances precisely? (Hint: think non-linear. You can make use of a software package such as ***Tablecurve***[1] to find other relationships.)

[1.3] Repeat Challenge [1.2] for an extrasolar planetary system. The data can be found in Table 16.2 in Milone and Wilson (2014).

[1.4] In Fig. 1.2, the Earth is located near the bottom of the page. Given that the plot represents the position of the Earth and Sun just before the June solstice, where would the Earth be located on its orbit at (a) the December solstice, and (b) the March equinox?

References

Bode, J.E.: (Deutliche) *Anleitung zur Kenntniss des gestirnten Himmels*, 2nd edition, pp. 462–463. Dieterich Anton Harmsen, Hamburg (1772). Cited in Jaki 1972

Cameron, A.G.W. The Primitive Solar Accretion Disk and the Formation of the Planets, In: S. F. Dermott (ed.), *The Origin of the Solar System*, 49–74. John Wiley, & Sons New York (1978)

Campbell, J.: *Historical Atlas of World Mythology* Vol. I: The Way of the Animal Powers. Harper & Row, New York (1988)

Chamberlain, T. C. *Carnegie Inst. Washington Yearbook*, 3, 133 (1904)

Chladni, E.F.F. *Über den Ursprung der von Pallas Gefundenen und anderer ihr ähnlicher Eisenmassen.* J. F. Hartknoch (ed.) (Riga: Repr. by Meteoritic Society, 1974), 56 (1794)

Hayashi, C., Nakazawa, K., Nakagawa, Y. Formation of The Solar System, In: Black D.C., Matthews, M.S. (eds.) *Protostars and Planets II*. University of Arizona Press, Tucson p. 1100–1153 (1985)

Jaki, S.L.: The Early History of the Titius-Bode Law. *Am. J. Phys.* **40**, 1014–1023 (1972)

Kelley, D.H., Milone, E.F.: *Exploring Ancient Skies*, 2nd edition. Springer Verlag, New York (2011)

Kepler, J. *Mysterium Cosmographicum*, 1981 translation by Duncan, A.M., Aiton, E.J. (eds.) Abaris Books, New York (1596)

Kuiper, G.P.: On the origin of the solar system. In: Hynek, J.A. (ed.) *Astrophysics*, pp. 365–424. McGraw Hill, New York (1951)

Marshack, A.: *The Roots of Civilization*. McGraw Hill, New York (1972)

Milone, E.F., Wilson, W.J.F.: *Solar System Astrophysics: Planetary Atmospheres and the Outer Solar System*, 2nd editon. Springer, New York (2014)

[1] Tablecurve 2D Automated Curve Fitting Software v2.0 1994 ed., published by Systat Software Incorporated. The current version 5.01 is available from Systat Software, Inc at: http://www.sigmaplot.com/products/tablecurve2d/

Moulton, F.R.: Evolution of the solar system. *Astrophys. J.* **22**, 166 (1905)

Nicholson, M.H.: *The Breaking of the Circle: Studies in the Effect of the "New Science" upon Seventeenth Century Poetry*. Northwestern University Press, Evanston (1950)

Nieto, M.M.: *The Titius-Bode Law of Planetary Distances*. Oxford University Press, Oxford (1972)

Perryman, M.: *The Exoplanet Handbook*, pp. 299–300. Cambridge University Press, Cambridge (2011)

Proctor, R.A.: *Other Worlds than Ours*, p. 220. D. Appleton, New York (1898)

Russell, H.N.: *The Solar System and its Origin*, pp. 95–96. Macmillan, New York (1935)

Safronov, V.S. *Evolution of the Protoplanetary Cloud and Formation* (Moscow: Nauka) (Tr. for NASA and NSF by Israel Prog. for Scientific Translations) (1969)

Thom, A.: *The Megalithic Lunar Observatories of Britain*. Oxford University Press, Oxford (1972)

von Weizsäcker, C.F.: Über die Entstehung des Planetensystems. *Zeitschrift für Astrophysik* **22**, 319 (1943)

Williams, J.P., Cieza, L.A.: Protoplanetary disks and their evolution. *Ann. Rev. Astronomy. and Astrophys.* **49**, 67–117 (2011)

Wurm, J.F.: *Astronomisches Jahrbuch fur das Jahr 1790*, pp. 161–171. G. J. Decker, Berlin (1787). Cited in Jaki 1972

Chapter 2
Basic Tools and Concepts

In this chapter, from the Greeks (through much subsequent development), we derive the tools of spherical astronomy. We will describe the basic theorems of spherical trigonometry and emphasize the usefulness of the sine and cosine laws. We will also describe the ellipse and its properties, in preparation for a subsequent discussion of orbits.

2.1 Circular Arcs and Spherical Astronomy

All astronomical objects outside the solar system are sufficiently far away that their shifts in position due to parallax (caused by periodic motions of the Earth) and proper motion (caused mainly though not exclusively by the objects' own motion) are too small to be discerned—at least by the unaided eye. Historically, this suggested that these objects could be regarded as being fixed to the inner surface of a sphere (or fixed to the outer surface of a transparent sphere), the *celestial sphere*, of some very large radius, centered on the Earth. Objects within the solar system change position with time, e.g., a superior planet's orbital motion causes an eastward motion across our sky relative to the distant stars and the Earth's orbital motion causes the planet to follow a retrograde loop. However, objects within the solar system can be referenced to the celestial sphere at any given instant of time.

We may wish to calculate the distance measured across the sky from one object to another knowing the distance of each of them from a third object, and also knowing an appropriate angle. ("Distance across the sky" is actually an arc length, measured in units of angle such as degrees or radians.) In doing this, we are in essence drawing arcs joining three objects to form a triangle on a spherical surface, so the mathematical relationships involved are those of spherical trigonometry. When we apply them to the sky we are practicing *spherical astronomy*. We note also that the objects do not need to be real; one or more of them can be a reference point, such as the north or south celestial pole.

E.F. Milone and W.J.F. Wilson, *Solar System Astrophysics: Background Science and the Inner Solar System*, Astronomy and Astrophysics Library,
DOI 10.1007/978-1-4614-8848-4_2, © Springer Science+Business Media New York 2014

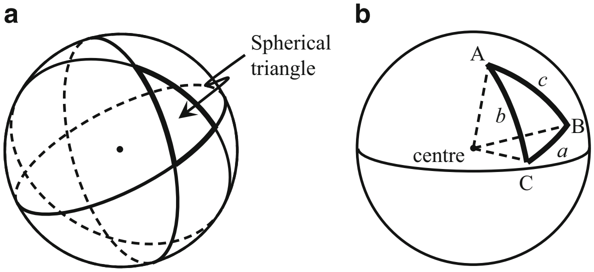


Fig. 2.1 The sphere with a spherical triangle on its surface

A *spherical triangle* is a triangle on the surface of a sphere such that each side of the triangle is part of a *great circle*, which has as its center the center of the sphere (a *small circle* will have its center along a radius of the sphere). For example, the Earth's equator is a great circle (assuming a spherical Earth), and any line of latitude other than the equator is a small circle.

The procedure in spherical astronomy lies primarily in the calculation of one side or angle in a spherical triangle where three other appropriate quantities are known, e.g., we may wish to find the length (in degrees) of one side of a spherical triangle given two other sides and the angle formed between them or find one side given a second side and the angle opposite each side or find one angle given a second angle and the side opposite each angle.

First, we review the basics of spherical trigonometry and then derive the cosine and sine laws, analogous but not identical to the cosine and sine laws of plane geometry. Additional theorems relating the three angles and three sides (involving, for example, haversines[1]) can also be found, but will not be derived here; for such theorems, see Smart's (1977) or Green's (1985) spherical astronomy texts, for example.

Figure 2.1a shows an example of a spherical triangle. The three ellipses are great circles seen in projection, and the spherical triangle (marked by heavy lines) is formed by their intersections. In Fig. 2.1b, we label the three sides of the triangle a, b, and c, and the angles at the three corners A, B, and C. We also draw a radius from the center of the sphere to each corner of the triangle (dashed lines).

Figure 2.2a shows an enlarged view of the spherical triangle and the radii to the center of the sphere. A circle is coplanar with its radii, so the shaded cross-section in this figure is a plane triangle (i.e., with straight sides).

It is important to distinguish between angles A, B, and C, which are the angles seen by an (two-dimensional) observer on the surface of the sphere and angles a, b, and c, which are the angles seen at the center of the sphere (Fig. 2.2b). That is, angle A is the angle (measured in degrees or radians) at the intersection of sides b and c on the surface of the sphere, whereas angle a is the angular separation "across the sky" from point B to point C (also measured in degrees or radians) as viewed from the center of the sphere. The principal equations relating these quantities to each other are the cosine and sine laws of spherical astronomy, which we will now derive.

[1] hav $\theta = (1/2)[1 - \cos\theta] = \sin^2(\theta/2)$.

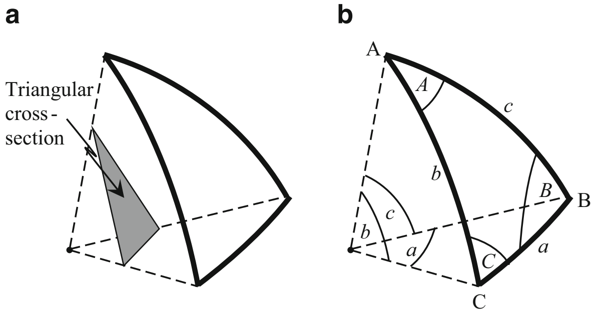


Fig. 2.2 Relating the sides of a spherical triangle to angles *a*, *b*, and *c* at the center of the sphere and angles *A*, *B*, and *C* on the surface of the sphere. The *heavy lines* are arcs of great circles; the *dashed lines* are radii of the sphere

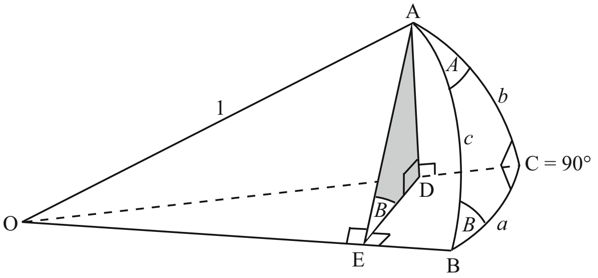


Fig. 2.3 Projection of a right spherical triangle onto a plane perpendicular to the base plane OBC. For convenience, we define points A and B to be located at the apices of angles *A* and *B*, respectively

NB: Small circle arcs have a different relation to interior angles. The laws derived here, therefore, do not apply to them.

2.1.1 The Law of Cosines for a Spherical Triangle

In this section we demonstrate a proof of the cosine law:

$$\cos a = \cos b\ \cos c + \sin b\ \sin c\ \cos A$$

First, we derive some useful results for a right spherical triangle ($\angle C = 90°$ in Fig. 2.3), which will also be useful in proving the sine law.

Take the radius of the sphere to be one unit of distance (OA = 1).

Note that any three points define a flat plane, so planes OAC, OAB, and OBC are each flat, and plane OAC is perpendicular to plane OBC. (To visualize this, it may help to think of arc BC as lying along the equator, in which case arc AC is part of a circle of longitude and arc AB is a "diagonal" great circle arc joining the two. Circles of longitude meet the equator at right angles, and longitudinal planes are perpendicular to the equatorial plane.)

From A, drop a line AD perpendicular to the plane OBC. Plane OAC ⊥ plane OBC, so point D is on the line OC. From D, draw a line DE ⊥ OB. (DE is, of course, not ⊥ OC.) Then:

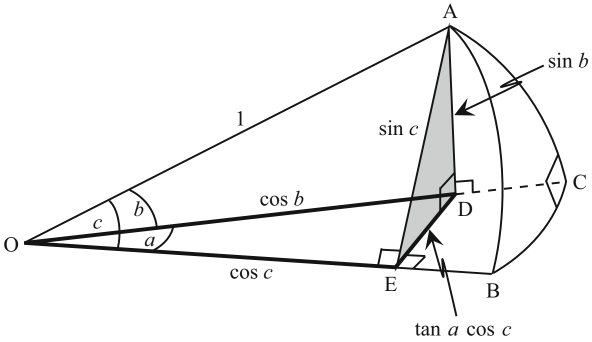


Fig. 2.4 Lengths of relevant sides, taking the radius of the sphere to be one unit of distance (OA = 1)

Plane ADE ⊥ plane OBC because line AD ⊥ plane OBC

Plane ADE ⊥ line OB because DE ⊥ OB and AD ⊥ plane OBC

∠ADE = 90° because plane OAC ⊥ plane OBC

∠OED = 90° and ∠OEA = 90° because plane ADE ⊥ line OB

Now define a plane tangent to the sphere (i.e., ⊥ to line OB) at point B; then ∠*B* lies in this tangent plane and, in fact, is equal to the angle of intersection of the two planes OAB and OBC. Plane ADE is parallel to the tangent plane, since both planes are ⊥ line OB, so it follows that:

$$\angle \mathrm{AED} = \angle \mathrm{B}$$

Then

$$\sin c = \frac{\mathrm{AE}}{\mathrm{OA}} = \frac{\mathrm{AE}}{1} = \mathrm{AE}$$
$$\cos c = \frac{\mathrm{OE}}{\mathrm{OA}} = \mathrm{OE}$$
$$\sin b = \frac{\mathrm{AD}}{\mathrm{OA}} = \mathrm{AD}$$
$$\cos b = \frac{\mathrm{OD}}{\mathrm{OA}} = \mathrm{OD}$$

These sides are shown in Fig. 2.4. Then from ΔODE,

$$\cos a = \frac{\mathrm{OE}}{\mathrm{OD}} = \frac{\cos c}{\cos b}$$

or

$$\cos c = \cos a \ \cos b \tag{2.1}$$

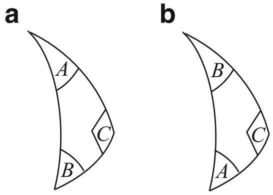


Fig. 2.5 Un-handedness of spherical triangles

Also from ΔODE,

$$\tan a = \frac{\mathrm{DE}}{\mathrm{OE}} = \frac{\mathrm{DE}}{\cos c}$$

or

$$\mathrm{DE} = \tan a \ \cos c$$

Therefore, from ΔADE,

$$\cos B = \frac{\mathrm{DE}}{\mathrm{EA}} = \frac{\tan a \ \cos c}{\sin c} = \tan a \ \cot c$$

Now $\angle C$ has been defined to be a right angle, but there is nothing to distinguish $\angle A$ from $\angle B$. It follows that any rule derived for the left-hand triangle in Fig. 2.5, above, has to be equally true for the right-hand triangle.

Thus, any rule derived for $\angle B$ is equally true for $\angle A$ with suitable relettering:

$$\cos A = \tan b \ \cot c \tag{2.2}$$

(One could derive this directly by redrawing Fig. 2.3 with the plane passing through point B perpendicular to the line OA instead of through point A perpendicular to the line OB.) Another equation can be obtained easily from ΔADE in Fig. 2.4:

$$\sin B = \frac{\mathrm{AD}}{\mathrm{AE}} = \frac{\sin b}{\sin c}$$

$$\therefore \sin b = \sin c \sin B \tag{2.3}$$

With (2.1) to (2.3) in mind, we can now look at the general spherical triangle (no right angles), as shown in Fig. 2.6.

Drop an arc $h \perp$ arc AB from point C to point D in Fig. 2.6. This divides the triangle into two right spherical triangles.

Define arc AD to take up an angle ϕ as seen from the center of the sphere (point O in Fig. 2.4). Then arc DB takes up angle $(c - \phi)$.

Apply (2.1) to each right spherical triangle in Fig. 2.6; then,

ΔADC: side b is opposite to the right angle, so,

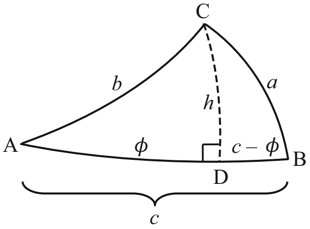


Fig. 2.6 The general spherical triangle as a combination of two right spherical triangles

$$\cos b = \cos h \cos \varphi \tag{2.4}$$

ΔBDC: side a is opposite to the right angle, so,

$$\cos a = \cos h \cos (c - \phi) \tag{2.5}$$

Divide (2.5) by (2.4) and use the standard trigonometric identity, $\cos(c - \phi) = \cos c \cos \phi + \sin c \sin \phi$, to get

$$\frac{\cos a}{\cos b} = \frac{\cos h \cos (c - \phi)}{\cos h \cos \phi} = \frac{\cos (c - \phi)}{\cos \phi}$$
$$= \frac{\cos c \cos \phi + \sin c \sin \phi}{\cos \phi} = \cos c + \sin c \tan \phi$$

or

$$\cos a = \cos b \cos c + \cos b \sin c \tan \phi \tag{2.6}$$

Now use (2.2) to obtain an equation for $\tan \phi$:

$$\cos A = \tan \phi \cot b = \frac{\tan \phi \cos b}{\sin b}$$

or

$$\tan \phi = \frac{\cos A \sin b}{\cos b} \tag{2.7}$$

Substituting (2.7) into (2.6), we arrive at,

$$\cos a = \cos b \cos c + \sin b \sin c \cos A \tag{2.8}$$

Equation (2.8) is the cosine law for spherical triangles. Because none of the angles in the triangle are right angles, there is nothing to distinguish one angle from another, and the same equation has to apply equally to all three angles:

$$\begin{aligned}\cos a &= \cos b\ \cos c + \sin b\ \sin c\ \cos A\\ \cos b &= \cos c\ \cos a + \sin c\ \sin a\ \cos B\\ \cos c &= \cos a\ \cos b + \sin a\ \sin b\ \cos C\end{aligned}$$

Note the canonical rotation of angles from one formula to the next.

2.1.2 Law of Sines for a Spherical Triangle

Application of (2.3) to ΔADC and ΔBDC in Fig. 2.6 gives, respectively,

$$\begin{aligned}\sin h &= \sin b\ \sin A \quad \text{and}\\ \sin h &= \sin a\ \sin B\end{aligned}$$

where sides b and a in Fig. 2.6 are opposite the right angles, and so replace side c in (2.3). Therefore,

$$\sin b \quad \sin A = \sin a\ \sin B \quad \text{and}$$
$$\frac{\sin a}{\sin A} = \frac{\sin b}{\sin B}$$

Again there is nothing to distinguish one angle from another, so this equation also has to apply equally to all angles:

$$\frac{\sin a}{\sin A} = \frac{\sin b}{\sin B} = \frac{\sin c}{\sin C} \tag{2.9}$$

Equation (2.9) is the law of sines for spherical triangles.

2.1.3 Other Laws

Two other formulae which may be useful in particular cases are the *analogue formula* of Smart (1977, p. 10):

$$\sin a\ \cos B = \cos b \sin c - \sin b\ \cos c\ \cos A \tag{2.10}$$

and Smart's (1977, p. 12) *four-parts formula*,

$$\cos a\ \cos C = \sin a\ \cot b - \sin C\ \cot B \tag{2.11}$$

The quantities in each of these may be canonically rotated, as per the cosine and sine laws.

2.1.4 Applications

Uses for spherical trigonometry abound. One example is to use a terrestrial system triangle to find the length of a great circle route for a ship or plane given the initial and final points of the route. The terrestrial coordinate system, (λ, ϕ), consists of latitude, ϕ, longitude, λ, the equator, the poles, and the sense in which latitude and longitude are measured (N or S from the equator and E or W from Greenwich, the "prime meridian," resp.). The units of longitude may be in expressed in units of hours, minutes and seconds of time or in units of degrees, minutes and seconds of arc; latitude is always given in units of arc. The two longitude arcs from the north or south pole to the initial and final points then form two sides of a spherical triangle, and the great circle route forms the third side.

A spherical triangle also can be used to find the arc length between two objects in the sky with a coordinate system appropriate to the sky, or to transform between two coordinate systems. Several different coordinate systems are in use. In the *horizon* or *altazimuth* (A, h) system, the coordinates are *altitude*, h, measured along a vertical circle positive toward the zenith from the horizon, and *azimuth*, A, measured from a fixed point on the horizon, traditionally the North point, CW around toward the East. Both are measured in degrees (and subunits) of arc. For example, an observer can use a theodolite to observe the altitude and azimuth of a star, then use these to find the latitude of the observing site.

For astronomical applications, corrections need to be made for the effect of the Earth's atmosphere on the altitude: refraction by the atmosphere raises the altitude, h, by a value which depends itself on the altitude. At the horizon, the correction is large and typically amounts to ~34′. At altitudes above about 40°, and expressed in terms of the zenith distance ($\zeta = 90^\circ - h$), the difference in altitude is ~57.3″ tan ζ. A measured altitude must be decreased by this amount to obtain the value of the altitude in the absence of the atmosphere. Another correction must be made for the "dip" of the horizon when the observer is not at ground or sea level (for example, when the observer is on the bridge of a naval vessel).

The *equatorial system* of astronomical coordinates has two variants:

The (H or HA, δ) system, which uses *Hour Angle*, H, and *declination*, δ. The (α, δ) system, which uses *Right Ascension*, α, and declination.

Thus, the location of astronomical objects such as the Sun or stars in the sky depends on the observer's latitude, the declination (distance above the celestial equator), δ, of the object, and the time of day.

Declination is measured North (+) or South (–) toward the N or S Celestial Poles from the *celestial equator*, the extension of the Earth's equator into the sky. Both H and α are measured in units of time. H is measured positive westward from the observer's meridian; α is measured eastward from the Vernal Equinox (the ascending node of the *ecliptic*; see below). Note that the (H, δ) system is dependent on the site, because H at any instant depends on the observer's longitude; the (α, δ) system is essentially independent of the observer's location. The latter is of use, for example, for a catalog of stars or other relatively 'fixed' objects.

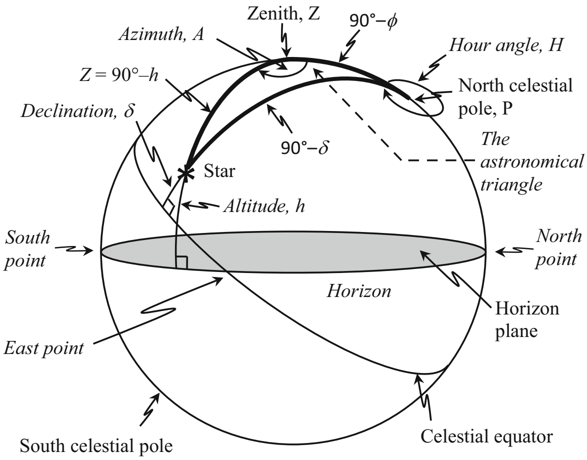


Fig. 2.7 The horizon A, h and equatorial H, δ coordinate systems. The *spherical triangle* whose apices lie at the star, the zenith, and the visible celestial pole is referred to as *the astronomical triangle*. Angle Z is the zenith angle of the star. Note that H, at the north celestial pole, is measured CW from south whereas A, at the zenith, is measured CW from north

At a particular site, H increases with time (the hour angle of the Sun + 12^h defines the apparent solar time) and this causes both altitude and azimuth to change with time also. Thus, the equatorial systems are more fundamental coordinate systems for celestial objects than the altazimuth system. The connection between the two variants of the equatorial system is the sidereal time, Θ:

$$\Theta = H + \alpha \tag{2.12}$$

The origin of the (α, δ) equatorial system is the Vernal Equinox, symbolized by the sign of Aries, γ, so the right ascension of this point is 0. Therefore sidereal time may be defined as,

$$\Theta \equiv H(\gamma) \tag{2.13}$$

(the hour angle of the Vernal Equinox). Recall that the Sun does not move along the celestial equator, but along the ecliptic, causing the different durations of sunshine with season (and latitude). Other consequences of this motion are the *equation of time* and the amplitude (maximum variation of the azimuth of the rising/setting Sun from the East/West points, respectively). See Fig. 2.7 for illustrations of the (A, h) and (H, δ) systems and the quantities needed to compute one set of coordinates from the other. In this figure, ϕ is the observer's latitude, which is equal to the altitude of the celestial pole above the observer's horizon and to the declination of the observer's zenith.

The *ecliptic coordinate system* (λ, β), involves the coordinates *celestial* (or *ecliptic*) *latitude*, β, and *celestial* (or *ecliptic*) *longitude*, λ, analogous to both the terrestrial coordinates (λ, ϕ) and the equatorial system (α, δ). The closer analogy, despite the names, is to the latter because the celestial longitude is measured CCW (viewed from the N) and from the same zero point, the Vernal Equinox, γ. The two

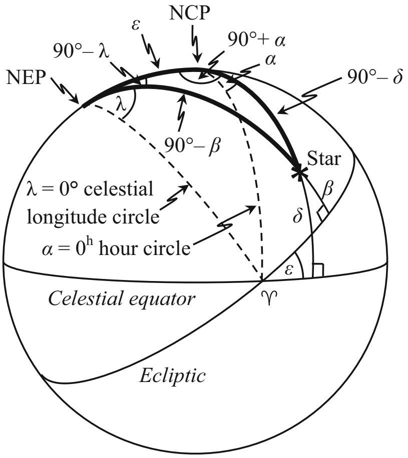


Fig. 2.8 The (λ, β) ecliptic and (α, δ) equatorial coordinate systems. ε is the obliquity of the ecliptic, i.e., the angle between the ecliptic and the celestial equator, and therefore also between the north ecliptic pole (NEP) and the north celestial pole (NCP). The spherical triangle involving the star, the NEP, and the NCP is used to obtain the transformation equations between the systems

reference circles, the celestial equator and the ecliptic, intersect at the vernal and autumnal equinoxes. The angle between them, known as the *obliquity of the ecliptic*, ε, is about 23.440° at present—it is slowly decreasing with time. The ecliptic system is very important for solar system studies and for celestial mechanics, both of which deal primarily with the solar system, the overlap with stellar kinematics, and dynamics notwithstanding. The relationship between the ecliptic and equatorial systems can be seen in Fig. 2.8, which shows the angles needed to compute one set of coordinates given the other.

The transformation equations between systems are readily obtained by drawing both on a celestial sphere and solving the resulting spherical triangles for the unknown pair of coordinates. Thus to compute the ecliptic longitude and latitude, draw the equatorial (RA) and ecliptic systems on the celestial sphere and use the separation of the poles of the system as one of the triangle legs. For this purpose, the spherical sine and cosine laws are perfectly adequate, even for checking the quadrant of the longitudinal coordinate (which may be in any of the four quadrants).

In Fig. 2.8, note that the great circle arc joining the NEP and the NCP is perpendicular to both the hour circle and the ecliptic/celestial longitude circle through the vernal equinox. The same arc is equal to the obliquity of the ecliptic, ε, the angle at the vernal equinox between the celestial equator and the ecliptic; Challenge [2.8(a)] at the end of this chapter invites you to prove that this is the case.

Example 2.1

An astronomer wants to observe a particular star with a telescope on an altazimuth mount. A star atlas provides the star's α and δ; and the star's hour angle is then given by $H = \Theta - \alpha$ from (2.12), where Θ is the sidereal time. However, because the mount is altazimuth, the coordinates actually needed are the altitude and azimuth. Find the transformation equations to convert the star's coordinates from H and δ to A and h. (Note how H and A are defined, and be careful with signs in equations containing trigonometric functions.) {Hint: The results given in Challenge [2.8] may be helpful.}

Solution to Example 2.1

Figure 2.7 shows the two relevant systems of coordinates. Here, ϕ is the observer's latitude and is equal to the declination of the observer's zenith. The star, the zenith, and the NCP form a spherical triangle referred to as the *astronomical triangle*, with sides $90° - h$ opposite the angle $360° - H$; $90° - \delta$ opposite the azimuth angle, A; and $90° - \phi$ opposite the angle formed at the star. Equations (2.8) and (2.9) then give, respectively,

$$\begin{aligned}\cos(90° - h) = {} & \cos\left(90° - \phi\right)\ \cos\left(90° - \delta\right) \\ & + \sin(90° - \phi)\ \sin\left(90° - \delta\right)\ \cos\left(360° - H\right) \\ \sin A = {} & \sin(90° - \delta)\ \sin(360° - H)/\sin(90° - h).\end{aligned}$$

The identities,

$$\begin{aligned}\sin(90° - \theta) &= \cos\theta \\ \cos(90° - \theta) &= \sin\theta \\ \sin(360° - \theta) &= -\sin\theta \\ \cos(360° - \theta) &= \cos\theta\end{aligned}$$

then give the transformation equations,

$$\sin h = \sin\phi\ \sin\delta + \cos\phi\ \cos\delta\ \cos H \tag{2.14}$$

$$\sin A = -\cos\phi\ \sin H/\cos h \tag{2.15}$$

Example 2.2

At some time of night, an observer at latitude 30° N sees a star at an altitude of 20° and an azimuth of 150°. Find the star's (H, δ) equatorial coordinates.

Solution to Example 2.2

Now we need the equations for the inverse of the transformation in Example 2.1, i.e., from the horizon to the equatorial system. Using Fig. 2.7 and the procedure in Example 2.1 again these are, from the cosine law,

$$\sin\delta = \sin\phi\ \sin h + \cos\phi\ \cos h\ \cos A \tag{2.16}$$

and, from the sine law,

$$\sin H = -\cos h\ \sin A/\cos\delta \tag{2.17}$$

Then with the values $\phi = 30°$, $h = 20°$, and $A = 150°$, (2.16) yields

$$\sin\delta = -0.53376, \text{so that}\, \delta = -32\overset{\circ}{.}260$$

and from (2.17), $\sin(H) = -0.55562$, so

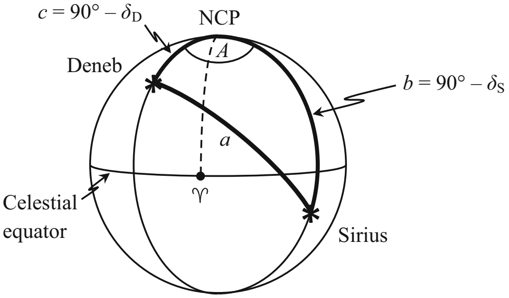


Fig. 2.9 The (α, δ) equatorial coordinate system, and the spherical triangle relating Deneb, Sirius, and the NCP. The RA arcs of the two stars are shown; angle A is the difference in right ascension between these two arcs

$$H = -33^{\circ}\!.753 \text{ or } -33^{\circ}\!.753/(15^{\circ}/\mathrm{h}) \cong -02^{\mathrm{h}}15^{\mathrm{m}} = 02^{\mathrm{h}}15^{\mathrm{m}}\mathrm{East}$$

Example 2.3

What is the angular distance across the sky from Deneb (α Cygni) at $\alpha = 20^{\mathrm{h}}\ 40^{\mathrm{m}}\ 24^{\mathrm{s}}$, $\delta = +45°\ 10'$ to Sirius (α Canis Majoris) at $\alpha = 6^{\mathrm{h}}\ 43^{\mathrm{m}}\ 48^{\mathrm{s}}$, $\delta = -16°\ 41'$?

Solution to Example 2.3

Figure 2.9 illustrates the (α, δ) equatorial coordinate system, the relevant spherical triangle, and the angles involved. The $\alpha = 0$ line is also shown for reference, as a dashed line from the NCP to the vernal equinox. Subscripts D and S signify Deneb and Sirius, respectively. We want to find the great circle arc length of side a.

The arc lengths of sides b and c are

$$b = 90^{\circ} - \delta_{\mathrm{S}} = 90^{\circ} - \left(-16^{\circ}\ 41^{'}\right) = 106^{\circ}\ 41^{'} = 10^{\circ}\!.668$$
$$c = 90^{\circ} - \delta_{\mathrm{D}} = 90^{\circ} - 45^{\circ}\ 10^{'} = 44^{\circ}\ 50^{'} = 44^{\circ}\!.83$$

There are several ways of expressing the angle A, all of which are equivalent by the fact that $\cos\theta = \cos(-\theta) = \cos(360°- \theta)$. Here we take the smallest positive value of A, but it would be equally correct and perhaps simpler to take $A = \alpha_{\mathrm{S}} - \alpha_{\mathrm{D}}$ to obtain a negative angle or $A = \alpha_{\mathrm{D}} - \alpha_{\mathrm{S}}$ to obtain a positive angle $>180°$.

$$c = 90^{\circ} - \delta_{\mathrm{D}}$$

From Fig. 2.9,

$$\begin{aligned} A &= \left(24^{\mathrm{h}} - \alpha_{\mathrm{D}}\right) + \alpha_{\mathrm{S}} = \left(24^{\mathrm{h}} - 20^{\mathrm{h}}40^{\mathrm{m}}24^{\mathrm{s}}\right) + 6^{\mathrm{h}}43^{\mathrm{m}}48^{\mathrm{s}} \\ &= 10^{\mathrm{h}}03^{\mathrm{m}}24^{\mathrm{s}} = \left(10^{\mathrm{h}} \times 15^{\circ}/\mathrm{h}\right) + \left(3^{\mathrm{m}} \times 1/60\,\mathrm{h/m} \times 15^{\circ}/\mathrm{h}\right) \\ &\quad + \left(24^{\mathrm{s}} \times 1/3600\,\mathrm{h/s} \times 15^{\circ}/\mathrm{h}\right) = 150^{\circ}\!.85 \end{aligned}$$

Then, using the cosine law to find side a,

$$\begin{aligned}\cos a &= \cos b\ \cos c + \sin b\ \sin c\ \cos A\\ &= \cos 106.68^\circ\ \cos 44.83^\circ + \sin 106.68^\circ\ \sin 44.83^\circ\ \cos 150.85^\circ\\ &= -0.7934.\end{aligned}$$

The inverse cosine is double-valued, so, within round-off error,

$$a = \cos^{-1}(-0.7934) = 142.6^\circ \text{ or } 217.4^\circ.$$

The shortest angular distance between any two objects on a sphere is always ≤180°, so the distance across the sky from Deneb to Sirius is $142^\circ\!.6$.

See Schlosser et al. (1991/1994) or Kelley and Milone (2011) for more details on and worked examples of transformations.

2.2 Properties of Ellipses

Spherical astronomy is a very important tool in solar system astronomy, as well as other areas, but it is not, of course, the only one. We now turn from a consideration of the circular and spherical to review basic properties of ellipses. This is useful for understanding the orbits of solar system objects, considered in Chap. 3, as well as extra-solar planets (Milone and Wilson 2014, Chap. 16).

An ellipse is the locus of points $P(x, \mathrm{y})$, the sum of whose distances from two fixed points is constant.

That is, in Fig. 2.10,

$$\boldsymbol{\ell}_1 + \boldsymbol{\ell}_2 = \text{constant} \tag{2.18}$$

The two fixed points are the *foci* of the ellipse (singular: *focus*). Relevant geometric definitions of the quantities in Figs. 2.10 and 2.11 are:

a = semi-major axis (therefore the length of the major axis is $2a$);
b = semi-minor axis (therefore the length of the minor axis is $2b$);
f = distance of each focus from the center of the ellipse;
r = distance from one focus to a point on the ellipse (e.g., point P);
r_{min} = distance from either focus to the nearest point on the ellipse;
$\mathrm{r}_{\mathrm{max}}$ = distance from either focus to the farthest point on the ellipse.

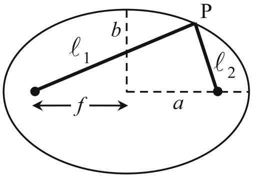


Fig. 2.10 The ellipse definition illustrated

Fig. 2.11 Major axis: $2a$

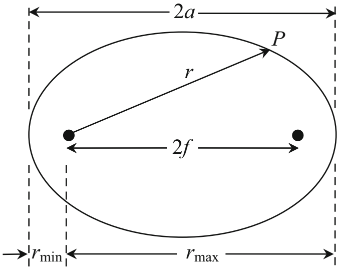
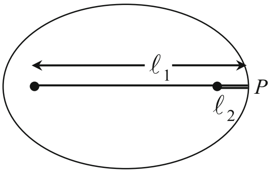


Fig. 2.12 Illustration of $\ell_1 + \ell_2 = 2a$



We can now evaluate the constant in (2.18), above. We do this by noting that (2.18) is true for every point on the ellipse. It is therefore also true if we move point P to the right end of the ellipse (Fig. 2.12) so that ℓ_1 and ℓ_2 lie along the major axis. The foci are symmetrically placed, so ℓ_2 equals the distance from the left focus to the left end of the major axis and ℓ_1 and ℓ_2 add up to $2a$ for this point.

Because $\ell_1 + \ell_2 =$ constant, independently of our choice of point P, we have

$$\boldsymbol{\ell}_1 + \boldsymbol{\ell}_2 = 2a \tag{2.19}$$

Some other relationships which follow from Figs. 2.10 and 2.11 are

$$\begin{aligned} r_{\min} &= a - f \\ r_{\max} &= a + f \end{aligned} \tag{2.20}$$

$$r_{\max} + r_{\min} = (a + f) + (a - f) = 2a \tag{2.21}$$

$$r_{\max} - r_{\min} = (a + f) - (a - f) = 2f \tag{2.22}$$

If we place point P at the end of the minor axis (Fig. 2.13), then $\ell_1 = \ell_2$ and it follows from $\ell_1 + \ell_2 = 2a$ that the distance from either focus to the end of the minor axis is equal to the length of the semi-major axis, a.

Then, using Pythagoras' theorem in Fig. 2.13, we have

$$b^2 = a^2 - f^2 = a^2\left(1 - \frac{f^2}{a^2}\right) \tag{2.23}$$

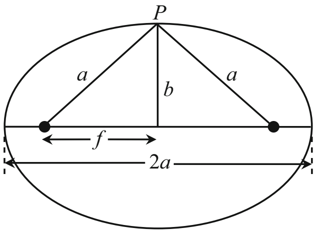


Fig. 2.13 The distance from either focus to one end of the semi-minor axis is equal to the length of the semi-major axis

Re-arranging (2.23) gives

$$f^2 = a^2 - b^2 \tag{2.24}$$

We now define the eccentricity, e, of the ellipse as the ratio of the distance between the foci to the length of the major axis:

$$e = \frac{2f}{2a} = \frac{f}{a} \tag{2.25}$$

Substitution of (2.25) into (2.23) gives

$$b^2 = a^2\left(1 - e^2\right) \tag{2.26}$$

Equations (2.20) and (2.25) then give

$$r_{\max} = a + f = a\left(1 + \frac{f}{a}\right) = a(1 + e) \tag{2.27}$$

$$r_{\min} = a - f = a\left(1 - \frac{f}{a}\right) = a(1 - e) \tag{2.28}$$

$$e = \frac{2f}{2a} = \frac{r_{\max} - r_{mim}}{r_{\max} + r_{mim}} \tag{2.29}$$

If we place the center of the ellipse at the origin of an (x,y) coordinate system as shown in Fig. 2.14, then we can express the equation of the ellipse in terms of x and y as follows. First, from Fig. 2.14 we have

$$\ell_1^2 = (f + x)^2 + y^2 \tag{2.30}$$

$$\ell_2^2 = (f - x)^2 + y^2 \tag{2.31}$$

We can now use (2.19), (2.24), (2.29) and (2.30) to eliminate ℓ_1, ℓ_2 and f and obtain a general equation for an ellipse in terms only of x, y, a and b. First, substitute the square roots of (2.29) and (2.30) into (2.19):

$$\sqrt{(f + x)^2 + y^2} + \sqrt{(f - x)^2 + y^2} = 2a \tag{2.32}$$

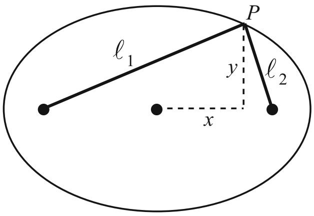


Fig. 2.14 *x* and *y* coordinates

then subtract the first term from both sides and square the result:

$$(f-x)^2+y^2=4a^2-4a\sqrt{(f+x)^2+y^2}+(f+x)^2+y^2 \tag{2.33}$$

Take all terms to the LHS except the term with the square root, and expand the squared terms in parentheses to obtain

$$fx+a^2=a\sqrt{f^2+2fx+x^2+y^2} \tag{2.34}$$

Square both sides:

$$f^2x^2+a^4=a^2f^2+a^2x^2+a^2y^2 \tag{2.35}$$

and substitute (2.24) into (2.33):

$$-b^2x^2=-a^2b^2+a^2y^2 \tag{2.36}$$

Finally, divide both sides by a^2b^2 and re-arrange to obtain the desired form of the equation for an ellipse:

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}=1 \tag{2.37}$$

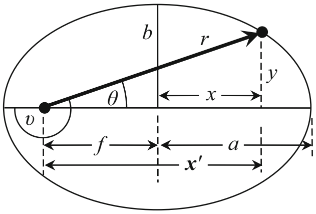
We can also write the equation for an ellipse in polar coordinates centered on one focus, as follows. From Fig. 2.15, substitute

$$\begin{aligned}&x=x^{'}-f=r\cos\theta-f\\&y=r\sin\theta\end{aligned}$$

into (2.37), square the numerators, and rearrange terms to obtain the quadratic equation

$$r^2\left(\frac{\cos^2\theta}{a^2}+\frac{\sin^2\theta}{b^2}\right)-r\left(\frac{2f\cos\theta}{a^2}\right)+\left(\frac{f^2}{a^2}-1\right)=0 \tag{2.38}$$

Fig. 2.15 Conversion from (*x*,*y*) to polar coordinates



The solution to (2.38) is

$$r = \frac{\dfrac{2f\cos\theta}{a^2} \pm \sqrt{\left(\dfrac{4f^2\cos^2\theta}{a^4}\right) - 4\left(\dfrac{\cos^2\theta}{a^2} + \dfrac{\sin^2\theta}{b^2}\right)\left(\dfrac{f^2}{a^2} - 1\right)}}{2\left(\dfrac{\cos^2\theta}{a^2} + \dfrac{\sin^2\theta}{b^2}\right)}$$

$$= \frac{f\cos\theta \pm a\sqrt{\cos^2\theta + \sin^2\theta\left(\dfrac{a^2 - f^2}{b^2}\right)}}{\cos^2\theta + \dfrac{a^2}{b^2}\sin^2\theta}$$

Now $a^2 - f^2 = b^2$, so the argument of the square root reduces to 1 and r becomes, with the help of (2.29) and (2.26),

$$r = \frac{f\cos\theta \pm a}{\cos^2\theta + \frac{a^2}{b^2}\sin^2\theta} = \frac{f\cos\theta \pm a}{\cos^2\theta + \frac{\sin^2\theta}{1-e^2}} = \frac{a\,(e\cos\theta \pm 1)}{\left(\frac{1-e^2\cos^2\theta}{1-e^2}\right)} \tag{2.39}$$

Regarding the ± sign, note that $e < 1$ and $\cos\,\theta \leq 1$. If we choose the negative sign then $(e\,\cos\,\theta - 1) < 0$ whereas the denominator is positive. This would make $r < 0$, which is physically impossible. Thus only the + sign is relevant. Equation (2.39) then reduces to,

$$r = \frac{a\,(1 - e^2)}{1 - e\cos\theta} = \frac{r_{\min}\,(1 + e)}{1 - e\cos\theta} \tag{2.40}$$

where $r_{\min} = a(1 - e)$ from (2.28). In celestial mechanics the customary angle used is the *argument of perihelion*, angle υ in Fig. 2.15; then $\cos\,\upsilon = \cos\,(\theta + 180°) = -\cos\,\theta$ and

$$r = \frac{a\,(1 - e^2)}{1 + e\cos\upsilon} = \frac{r_{\min}\,(1 + e)}{1 + e\cos\upsilon} \tag{2.41}$$

Table 2.1 Eccentricities of ellipses and related curves

Curve	Eccentricity
Circle	0
Ellipse	$0 < e < 1$
Parabola	1
Hyperbola	>1

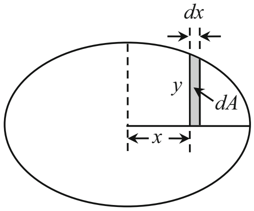


Fig. 2.16 Finding the area of an ellipse

The left equation of (2.40) describes an elliptical orbit. Kepler's first "law" is that planets move in elliptic orbits with the Sun at one focus. We now look at different limiting cases of the eccentricity. First, if we set $f = 0$ then the two foci coincide at the center of the ellipse and $e = f/a = 0$. Then (2.23) and (2.26) both give $b = a$ (i.e., the semi-major and semi-minor axes are equal), and from (2.37) we have

$$\frac{x^2}{a^2} + \frac{y^2}{a^2} = 1 \quad \text{or} \quad x^2 + y^2 = a^2,$$

which is the equation for a circle. Thus a circle is an ellipse with an eccentricity of zero. If, on the other hand, we let $f \to \infty$ at constant $r_{\min}$, then from (2.25) and (2.28),

$$e = \frac{f}{a} = \frac{f}{f + r_{\min}} \to 1$$

The resulting curve is a parabola. If $e > 1$ then we have a hyperbola. These results are summarized in Table 2.1.

The area of an ellipse can be found by integration, as intimated in Fig. 2.16:

$$A = 4\int_{x=0}^{a} dA = 4\int_{x=0}^{a} y\,\mathrm{d}x \tag{2.42}$$

where, by (2.37), we have

$$y = b\left(1 - \frac{x^2}{a^2}\right)^{\frac{1}{2}} \tag{2.43}$$

This can be integrated with a trigonometric substitution,

$$x = a \sin\theta \tag{2.44}$$

to obtain

$$A = \pi ab = \pi a^2\sqrt{1 - e^2} \tag{2.45}$$

Challenges

[2.1] Compute the great circle distance and the initial and final bearings (=azimuths) for a voyage to Singapore ($\lambda = 103{:}^\circ85$ E, $\phi = 1{:}^\circ28$ N) from Vancouver ($\lambda = 123{:}^\circ2$ W, $\phi = 49{:}^\circ3^\circ$N).

[2.2] Use spherical trigonometry to derive the equations of transformation between the equatorial and ecliptic systems. Figure 2.8 may be helpful.

[2.3] Compute the celestial longitude and latitude of an object at $\alpha = 15^\mathrm{h}\ 39^\mathrm{m}\ 40^\mathrm{s}$, $\delta = -5^\circ\ 17.8'$.

[2.4] Derive the equations of transformation between the horizon (A, h) and (H, δ) equatorial system to obtain the Hour Angle, H, and the declination, δ, in terms of the altitude, h, the azimuth, A, and the observer's latitude, ϕ.

[2.5] For a site with latitude $\lambda = 38^\circ$ N, what are the equatorial coordinates of:

(a) a star located at the NCP?
(b) a star on the horizon at the South point at 12^h local sidereal time?
(c) a star overhead at local midnight on March 21?

[2.6] For a site with latitude $\lambda = 38^\circ$S, what are the equatorial coordinates of:

(a) a star located at the SCP?
(b) a star on the horizon at the North point at 12^h local sidereal time?
(c) a star overhead at local midnight on March 21?

[2.7] Prove that the altitude of the North Celestial Pole in Fig. 2.7 is the latitude of the observer.

[2.8] Figure 2.17 shows certain aspects of Fig. 2.8. Prove that

(a) the great circle arc EC between the north ecliptic pole and the north celestial pole, labeled θ in Fig. 2.17, is equal to the obliquity of the ecliptic, ε; and
(b) angles x and y are both equal to 90°. [Hint: Both the sine and cosine laws may be helpful.]

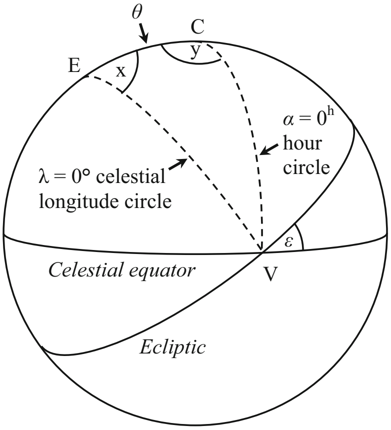


Fig. 2.17 Diagram for Challenge [2.8]. E = North Ecliptic Pole; C = North Celestial Pole; V = Vernal Equinox; ε = obliquity of the ecliptic; θ = arc EC

References

Green, R.M. *Spherical Astronomy*. Cambridge University Press, Cambridge 1985

Kelley, D.H., Milone, E.F. *Exploring Ancient Skies*, 2nd edn. Springer, New York 2011

Milone, E.F., Wilson, W.J.F. *Solar System Astrophysics: Planetary Atmospheres and the Outer Solar System*, 2nd edn. Springer, New York 2014

Schlosser, W., Schmidt-Kaler, T., Milone, E.F. *Challenges of Astronomy: Hands-On Experiments for the Sky and Laboratory*. Springer, New York (1991/1994)

Smart, W.M. *Spherical Astronomy*, 6th or later ed., revised by R. M. Green. Cambridge University Press, Cambridge 1997

Chapter 3
Celestial Mechanics

An important field for solar system studies is the area of physics of motion, that is, kinematics and dynamics, as applied to celestial objects. This field is called celestial mechanics. It is impossible to do full justice to this subject within a single chapter, but we will provide an introduction that may be useful for further studies and provide a background to understanding planetary dynamics as we apply it in this book. We begin first with an understanding of the orbits of the planets and of the other objects which move about the Sun. This will prove invaluable also in studies of extrasolar planets in orbit around their stars.

3.1 The Two-Body Problem

The geometry of the solar system can be considered on many scales. On the largest scales, the potential well created by the Sun dominates the scene; on the smallest gravitational scales (that we have discerned), the mutual interactions of moons and ring fragments are noticeable (see Chap. 13 of Milone and Wilson 2014). The basic equations in celestial mechanics are Newton's gravitational law and laws of motion.

If there were only a single mass in space, with no force acting on it, the motion would be unchanging, according to *Newton's first law of motion*: if at rest, it would stay at rest; if moving, its velocity would be constant in both magnitude and direction.

Given two objects of mass m_1 and m_2, the condition for Newton's first law of motion no longer holds. In this case, we want to determine the orbit of either one around the other, assuming that the only force acting is their mutual gravitational attraction. In the process, we will derive two of Kepler's three laws as well as a number of other useful quantities and equations.

In Fig. 3.1, point O marks the origin of an inertial (non-accelerating) reference frame, with masses m_1 and m_2 at positions marked by the vectors $\mathbf{r}_1$ and $\mathbf{r_2}$, respectively. We take the vector $\mathbf{r}$ as pointing from m_1 to m_2, and the unit vector, $\hat{\mathbf{r}}$, as pointing in the same direction (i.e., radially out from m_1 to m_2). Note that

E.F. Milone and W.J.F. Wilson, *Solar System Astrophysics: Background Science and the Inner Solar System*, Astronomy and Astrophysics Library,
DOI 10.1007/978-1-4614-8848-4_3, © Springer Science+Business Media New York 2014

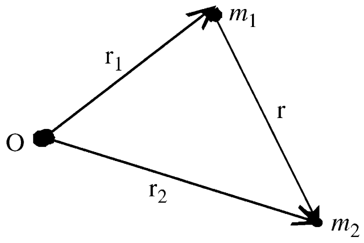


Fig. 3.1 Two masses in an inertial coordinate system

$$\mathbf{r} = r\hat{\mathbf{r}}$$

where the quantity r is a scalar, having size or magnitude, but containing no direction information.

According to Newton's third law of motion, the forces exerted by each object on the other are equal in magnitude and opposite in direction, and, according to Newton's law of gravitation, the magnitude of each force is

$$\mathrm{F}_g = \frac{Gm_1m_2}{r^2} \tag{3.1}$$

Then by Newton's second law of motion[1] the force on m_1 is

$$\mathbf{F}_1 = m_1\mathbf{a}_1 \equiv m_1\ddot{\mathbf{r}}_1 = \frac{Gm_1m_2}{r^2}\hat{\mathbf{r}} \tag{3.2}$$

Then the equation of motion of m_1 in the frame shown in Fig. 3.1 is given by

$$\ddot{\mathbf{r}}_1 = \frac{Gm_2}{r^2}\hat{\mathbf{r}} \tag{3.3}$$

Here we have shown time derivatives by dots over the quantity, so, for example, the acceleration, a, of m, can be written as "r double dot." Formally,

$$\dot{\mathbf{r}} \equiv \frac{d\mathbf{r}}{dt}; \qquad \ddot{\mathbf{r}} \equiv \frac{d^2\mathbf{r}}{dt^2}$$

[1] Newton's second law of motion: $\sum \vec{F} = m\vec{a}$. The acceleration, $\vec{a}$, of a body is proportional to the sum of the forces, $\sum \vec{F}$, acting on it and inversely proportional to its mass, m.

Similarly, the equation of motion of m_2 is obtained from

$$\mathbf{F}_2 = m_2\mathbf{a}_2 \equiv m_2\ddot{\mathbf{r}}_2 = \frac{Gm_1m_2}{r^2}(-\hat{\mathbf{r}}\,) \tag{3.4}$$

where $\mathbf{F}_2$ is in the negative or opposite direction to the vector $\mathbf{r}$. Then,

$$\ddot{\mathbf{r}}_2 = -\frac{Gm_1}{r^2}\hat{\mathbf{r}} \tag{3.5}$$

Now $\mathbf{r} = \mathbf{r_2} - \mathbf{r_1}$ as shown in Fig. 3.1. Differentiating twice with respect to time gives

$$\ddot{\mathbf{r}} = \ddot{\mathbf{r}}_2 - \ddot{\mathbf{r}}_1 \tag{3.6}$$

so that

$$\begin{aligned} \ddot{\mathbf{r}} &= \left(-\frac{Gm_1}{r^2}\hat{\mathbf{r}}\right) - \left(+\frac{Gm_2}{r^2}\hat{\mathbf{r}}\right) \qquad \text{or} \\ \ddot{\mathbf{r}} &= -\frac{G\,(m_1+m_2)}{r^2}\hat{\mathbf{r}} \end{aligned} \tag{3.7}$$

The $(-)$ sign in this equation shows that the acceleration acts to shorten $\mathbf{r}$, that is, gravity is an attractive force. This is Newton's gravitational law for relative orbits.

With (3.7), we can forget about $\mathbf{r_1}$ and $\mathbf{r_2}$ and work directly with $\mathbf{r}$. Consider the general expression for the force exerted on a planet of mass m, by the Sun, with mass $\mathfrak{M}_\odot$:

$$\mathbf{F} = -\left[(G\mathfrak{M}_\odot m)/r^2\right]\hat{\mathbf{r}} = -\left[(G\mathfrak{M}_\odot m)/r^3\right]\mathbf{r} \tag{3.8}$$

where the vector $\mathbf{r} = r\,\hat{\mathbf{r}}$ is in the direction of the planet (of mass m), away from the Sun and G is the gravitational constant[2] with the determined value $6.67384 \pm 0.00080 \times 10^{-11}$ m^3 kg^{-1} s^{-2}. Here, again, the negative sign indicates an attraction toward the Sun. Newton's second law of motion then permits us to arrive at the critical differential equation, equivalent to (3.7):

$$m\ddot{\mathbf{r}} = \; -\left[\left(G(\mathfrak{M}_\odot + m)/r^3\right]\mathbf{r}\right. \tag{3.9}$$

[2] The latest values for physical constants, generally accepted worldwide, may be found on the National Institute for Standards and Technology (NIST) website, http://physics.nist.gov/cuu/Constants/. At current writing (May 2013), these are the 2010 CODATA recommended values, where CODATA is the Committee on Data for Science and Technology.

By Newton's third law of motion, the force of the Sun on the planet is equal in magnitude to that of the planet on the Sun, so that $\mathfrak{M}_\odot \ddot{r}_\odot = -m\ddot{r}$. Thus the change in velocity of the Sun is very small compared to that of the planet. Nevertheless, an extra-solar planet may sometimes be detected by such very small accelerations in the motion of its star (see Sect. 16.2.1 in Milone and Wilson 2014).

For present purposes, we note that, in the relative orbit, (3.9) allows us to ignore the mass of the lesser object if it is small enough. For planets in our solar system, $m \ll \mathfrak{M}_\odot$, so we can set $\mathfrak{M}_\odot + m \approx \mathfrak{M}_\odot$, a condition that holds to one part in 1,000 or better.

In spherical coordinates (r, the distance from the Sun, and θ, the angular variable), one may express the velocity as: $\dot{\mathbf{r}} = \dot{r}\,\hat{\mathbf{r}} + r\dot{\theta}\,\hat{\theta}$, and making use of the conventions $\mathrm{D}\,\hat{\mathbf{r}} = \dot{\theta}\,\hat{\theta}$, and $\mathrm{D}\,\hat{\theta} = -\dot{\theta}\,\hat{\mathbf{r}}$, where $\hat{\theta}$ is the coordinate direction ⊥ to $\hat{\mathbf{r}}$, and D is the time-derivative operator, d/dt. Then, applying D to $\dot{\mathbf{r}}$, we get for the relative acceleration

$$\ddot{\mathbf{r}} = \left(\ddot{r} - r\dot{\theta}^2\right)\hat{\mathbf{r}} + \left(r\ddot{\theta} + 2\dot{r}\dot{\theta}\right)\hat{\theta} \tag{3.10}$$

By definition there is no $\hat{\theta}$ component for a centrally directed force, so $r\ddot{\theta} + 2\dot{r}\dot{\theta} = 0$. However, $r\ddot{\theta} + 2\dot{r}\dot{\theta} = (1/r)\,\mathrm{D}\left(r^2\dot{\theta}\right)$ so $\mathrm{D}\left(r^2\dot{\theta}\right) = 0$, and therefore

$$r^2\dot{\theta} = \text{constant} \tag{3.11}$$

The *constant* is usually called h, the *angular momentum* (per unit mass). Because

$$r^2\dot{\theta} = |\mathbf{r} \times \dot{\mathbf{r}}| = r\dot{r}\sin\phi \tag{3.11a}$$

where ϕ is the angle between the vectors $\mathbf{r}$ and $\dot{\mathbf{r}}$, it can be seen that h is actually a vector and is ⊥ to the plane containing $\mathbf{r}$ and $\dot{\mathbf{r}}$. The constancy of $\mathbf{h}$ in magnitude and in direction assures that the planet continues to move only in the original plane.

It may be also shown that the rate of change of the area swept out by $\mathbf{r}$, the areal velocity, is

$$\Delta A/\Delta t = r^2\dot{\theta}/2 = h/2 \tag{3.12}$$

which is a succinct statement of Kepler's second law.

To arrive at the energy equation for an orbit, form the dot product of the vectors $\dot{\mathbf{r}}$ and $\ddot{\mathbf{r}}$:

$$\dot{\mathbf{r}} \cdot \ddot{\mathbf{r}} = 1/2D(\dot{\mathbf{r}} \cdot \dot{\mathbf{r}}) = 1/2D\left[\dot{r}^2 + \left(\dot{\theta}\,r\right)^2\right] \tag{3.13}$$

and also, from (3.9), writing $\mu = G(\mathfrak{M}_\odot + m)$,

$$\dot{\mathbf{r}} \cdot \ddot{\mathbf{r}} = -\mu \dot{\mathbf{r}} \cdot \mathbf{r}/r^3 = -\mu \dot{r}/r^2 = D(\mu/r) \tag{3.14}$$

Combining the integrands of (3.13) and (3.14), we obtain the *orbital energy equation*:

$$(1/2)\left(\dot{r}^2 + r^2\dot{\theta}^2\right) - (\mu/r) = \text{constant} \tag{3.15}$$

where the constant is the total energy per unit mass, E. With (3.11), (3.15) becomes

$$\dot{r}^2 = 2\,(E + \mu/r) - h^2/r^2 \tag{3.16}$$

Making use of the equation

$$\mathrm{d}r/\mathrm{d}t = (\mathrm{d}r/\mathrm{d}\theta)\,(\mathrm{d}\theta/\mathrm{d}t) \tag{3.17}$$

and making use of (3.12) and (3.17) in (3.16), we get

$$\mathrm{d}\theta/\mathrm{d}r = \left(h/r^2\right)/\sqrt{2(E + \mu/r) - h^2/r^2} \tag{3.18}$$

After a change in variable, $u = 1/r$, recognizing that $\mathrm{d}r/r^2 = \mathrm{d}(-1/r)$, and dividing numerator and denominator by h, (3.18) becomes, in integral form,

$$\theta = -\int \left(a + bu - u^2\right)^{-1/2} du \tag{3.19}$$

where constants $a \equiv 2E/h^2$ and $b \equiv 2\ \mu/h^2$; N.B.: this a is not the semi-major axis of Sect. 2.2, nor is it the logarithmic base of the Blagg-Richardson formula of Chap. 1.

The solution, by quadratures, is found in tables of integrals, such as Peirce (1957), where it is No. 166:

$$\int \mathrm{d}x(X)^{-1/2} = -(-c)^{-1/2}\arcsin\left[(2cx + b)(-q)^{-1/2}\right] \tag{3.20}$$

where $X = a + bx + cx^2$, c $= -1$, and $q = 4ac - b^2$, so that the solution is

$$\theta = \arcsin\left[(-2u + b)\left(4a + b^2\right)^{-½}\right] + \gamma \tag{3.21}$$

and γ is a constant of integration. From this,

$$u = \left(\mu/h^2\right)\left\{1 - \left[2Eh^2/\mu^2 + 1\right]^{1/2} \sin\left(\theta - \gamma\right)\right\} \tag{3.22}$$

We take γ to be 90° so that $\sin(\theta - \gamma) = \sin(\theta - 90°) = -\cos\theta$; then

$$r = \left(h^2/\mu^2\right)/\left\{\left[\sqrt{\left(2Eh^2/\mu^2\right) + 1}\right]\cos\theta + 1\right\} \tag{3.23}$$

Setting

$$\begin{aligned} h^2/\mu = a(1 - e^2) = q\left(1 + e\right) \qquad &\text{and} \\ \sqrt{\left(2Eh^2/\mu^2\right) + 1} = e \end{aligned} \tag{3.24}$$

where q = a(1 − e) is the perihelion distance (see Sect. 2.2), a is the semi-major axis, and e is the eccentricity, (3.24) becomes

$$r = q(1 + e)/[e\cos\theta + 1] = q(1 + e)/[1 + e\cos\theta] \tag{3.25}$$

This form is applicable to all orbits, i.e., with all values of eccentricity, but ellipses are much more commonly encountered in the solar system.

3.2 Orbital Elements

That planets move in ellipses, is the statement of Kepler's first law.[3] The properties of the orbit can be summarized by the constants of the integration. The basic differential equation (3.10) is a second-order equation, so there are six constants of integration. We have discussed the quantities h and E, which are related to the geometrical quantities a and e. The geometrical orbital elements are:

a: the *semi-major axis* of the ellipse, which establishes the scale of the orbit
e: the *eccentricity*, which establishes the shape
i: the *inclination*, which establishes the tilt of the orbital plane, relative to the ecliptic plane
Ω: the *longitude of the ascending node* (ascending crossover point of the orbit, measured with respect to the line to the vernal equinox), which establishes the orientation of the orbital plane
ω: the *argument of perihelion* (measured from the ascending node to the perihelion point; the sense is in the direction of orbital motion), which establishes the orientation of the ellipse within the orbital plane
T_0: the *epoch*, which in this context, is an instant when the object is at perihelion

[3] More strictly, we should write Kepler's first "law" as we do in Chap. 1, because Kepler's "laws" are actually empirical approximations to the true motions of planets. Planetary orbits differ from ellipses because of the perturbing effects of the other planets and asteroids in the solar system, and the effects of General Relativity. However, we here follow common usage and forego the quotation marks.

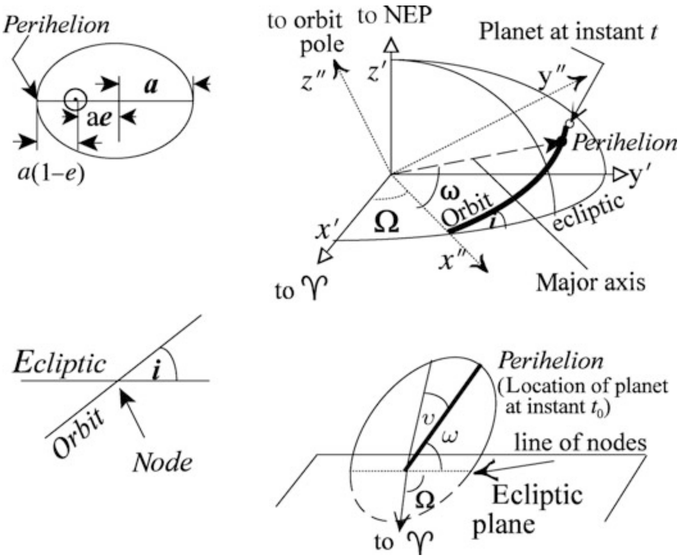


Fig. 3.2 Orbital elements (*bold*) and other quantities, in plan, elevation, and oblique perspectives of the planetary orbit and the ecliptic

The elements are illustrated in Fig. 3.2. If the inclination is so close to zero that the ascending node is not well-established, the *longitude of perihelion*, ϖ ("curly pi") $= \Omega + \omega$ may be used in place of Ω and ω. The *period* of revolution, P, is sometimes included in the elements; it is not quite redundant, even though Kepler's third law ($\mu P^2 = 4\pi^2 a^3$), would suggest that it is [recall that the quantity $\mu = G(\mathfrak{M}_\odot + m)$ really depends on the mass of the planet as well as that of the Sun].

In the computation of ephemerides (predictions of position of the planet with time), the angular variable (θ), or more specifically the *true anomaly*, (usually represented by the lower case Greek letter upsilon, υ), identical to θ in (3.23) and (3.25), is computed from more directly time-related quantities, the *eccentric anomaly* and the *mean anomaly*. Kepler's equation, discussed in Sect. 3.5, relates these quantities. In Sect. 3.4 we describe the computation of ephemerides given the orbital elements.

In general, nature is more complicated than our self-consistent discussion of orbits would lead one to believe. Although the motions of two isolated bodies can be solved analytically and therefore exactly, this is usually not true for more than two bodies. The reason: mutual gravitational interactions make exact solutions impossible. However, numerical methods and high-speed computation have made dramatic progress in n-body analyses and computation of ephemerides. In Sect. 3.8, we discuss the evaluation of perturbations to orbital motions. Moreover, there are exceptions to the lack of closed analytic solutions. In the solar system, for instance, the interaction of some asteroids with Jupiter and the Sun illustrates a famous condition: *the restricted three-body problem*. The condition is also approximated in the satellite systems of the giant planets. Therefore, we describe this condition next.

3.3 The Restricted Three-Body System

Under certain circumstances in a three-body system, a close approximation to an exact solution is possible. If two of the bodies, m_1 and m_2, are massive (e.g., the Sun and Jupiter) and the third body, m_3, has an infinitesimal and therefore negligible mass, and the orbits of m_1 and m_2 are circular, then an analytic solution can be found. Because real orbits will not be exactly circular, and no real body can be completely massless, only an approximation to this condition can be achieved. But the masses of the Sun and Jupiter are so large relative to the small bodies of the solar system, and the eccentricity of Jupiter's orbit sufficiently small, that the approximation is valid to rather high precision. The solutions include several points in a frame of reference that co-rotates with the two massive objects. At these *Lagrangian points*,[4] the third body (or group of small bodies) can, once there, remain. These are called L_1, ..., L_5, depending on their locations with respect to the greater and smaller of the two large masses. As illustrated in Fig. 3.3, created with the aid of Binary Maker 3 (Bradstreet and Steelman 2004), L_1 is between the two massive objects; the colinear points L_2 and L_3 are on either side, and all three are on the line joining the centers of the massive pair of objects. L_4 and L_5 are perpendicular to this line and form equilateral triangles with the massive objects.

L_1, L_2, and L_3 are situated on curves known as zero-velocity curves, two of which are shown in Fig. 3.3. These curves are the cross-sections in the plane of the orbit of potential surfaces, which confine the motions of m_3 if it is located on one of them. Objects located at L_1, L_2, and L_3 may be perturbed away from these points easily, but those located at L_4 and L_5 are stable against all but the strongest perturbations.

The surfaces, curves, and points are discussed in detail by Moulton (1914/1958, pp. 281ff), Danby (1988), pp. 255ff), and by Murray and Dermott (1999, pp. 77ff), among many sources.

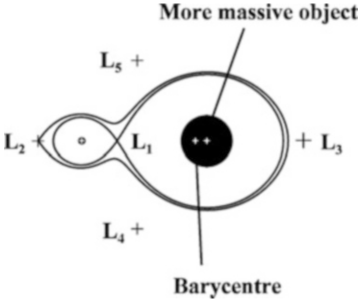


Fig. 3.3 The Lagrangian points: solutions of the restricted three-body problem. Illustration created with the help of Binary Maker 3 (Bradstreet and Steelman 2004). Two of the zero-velocity curves are shown also

[4] named after Joseph Louis Lagrange (1736–1813).

In the solar system, the planets revolve CCW as viewed from above the north ecliptic pole (NEP); thus with respect to the smaller of the two major masses, L_4 is on the leading side, and L_5 on the trailing side of its orbit.

To return to our example, the Greek and Trojan asteroids continuously chase each other around the solar system while Zeus (Jupiter) and Apollo (the Sun) dominate the scene and witness the events. In fact, there are asteroids named for Homer's Greeks and Trojans at both the leading L_4 and trailing L_5 points, but in general, smaller bodies located at either of the two Lagrangian points are called "Trojans." There are Trojan asteroids found around the orbits of Neptune, Saturn, Mars, and one is known around Earth's orbit. Moreover, there are even Trojan satellites in place around the giant planets (see Milone and Wilson 2014, Chap. 13).

Orbital parameters of planetary orbits vary with time, mostly in cyclic ways. The motions of the smaller bodies are generally more complex and may be unpredictable (aspects of Pluto's motion can be characterized this way), for reasons that we discuss later. For an instant of time, the orbital elements characterizing a particular orbit may be specified: they are the "osculating" orbit elements, i.e., they "kiss" the true orbit at the epoch for which the elements are specified. Beyond this instant, each element is subject to perturbations.

3.4 Computation of Ephemerides

If an instant of time is specified, then, given the orbital elements, the position in the orbit, and in the sky, can be obtained as follows.

1. Compute first the *mean anomaly M* from its defining equation:

$$M \equiv 2\pi(t - T_0)/P \tag{3.26}$$

   where t and T_0 are expressed in Julian Day numbers (JDN), typically. Note that M is an angular variable that varies linearly with time, so that if the orbit were circular, M would be identical with the true anomaly, υ. Usually, $e \neq 0$; but in this case, depending on the size of the eccentricity, υ can be found either (a) from an approximation in terms of M directly or (b) more exactly but less directly, with the help of the eccentric anomaly, E, which we define and describe in Sect. 3.5, below. For now, we take route (a); therefore, having M, we find υ as follows.
2. If the eccentricity is sufficiently small, by which we mean $e < 0.6623$ (if not, see the next section!), we may use the approximation (Moulton 1914/1958, p. 171),

$$\begin{aligned}\upsilon = M + 2e\sin M + (5/4)e^2\sin 2M + \left(1/12\right)e^3\left(13\sin 3M - 3\sin M\right)\\ +(1/96)e^4\left(103\sin 4M - 44\sin 2M\right) + \ldots\end{aligned} \tag{3.27}$$

   Then, with υ in hand, with (3.25), and the substitution $q = \mathrm{a}(1 - \mathrm{e})$,

3. *Compute the radius vector:*

$$r = a(1 - e)^2/(1 + e\ \cos \upsilon) \tag{3.28}$$

The rectangular (right hand system) coordinates in the orbital plane are then $x_0 = r\ \cos\ \upsilon$, $y_0 = r\ \sin\ \upsilon$, $z_0 = 0$, where the x-axis is in the direction of perihelion (more generally, *pericenter*), and the y-axis, 90° ahead in the plane of the orbit. This gives the position of the planet within the orbit.

As an aside, in addition to the variable υ, the true or mean longitudes are sometimes used to describe the location of an object in the orbital plane. These are defined as follows:

- *True longitude*, $L = \Omega + \omega + \upsilon$. In turn the *longitude of perihelion*, $\varpi = \Omega + \omega$ is sometimes used, especially when the node is ill-defined (i.e., when the inclination is near zero), and the *argument of latitude*, $u = \omega + \upsilon$ is sometimes used, especially when the pericenter is ill-defined because e is very close to zero.

  Thus, $L = \varpi + \upsilon$ and $L = \Omega + u$ are equivalent.
- *Mean longitude*, $\ell = \varpi + n(t - T_0) = \varpi + M$ where n is the mean motion, formally defined by (3.39) below, and M is the mean anomaly.

4. To specify where the planet is in the sky, we can first calculate the position in the ecliptic and then transform this into the equatorial system (see Sect. 2.1.4) or we may make use of a series of auxiliary relations, first introduced by Karl Friedrich Gauss (1777–1855) to go more directly into the equatorial system (see Moulton 1914/1958, p. 188):

$$\begin{aligned}
\alpha' \sin A &= \cos\ \Omega \\
\alpha' \cos A &= -\sin\ \Omega\ \cos\ i \\
\beta' \sin B &= \sin\ \Omega\ \sin\ \varepsilon \\
\beta' \cos B &= \cos\ \Omega\ \cos\ i\ \cos\ \varepsilon - \sin\ i\ \sin\ \varepsilon \\
\gamma' \sin C &= \sin\ \Omega\ \sin\ \varepsilon \\
\gamma' \cos C &= \cos\ \Omega\ \cos\ i\ \sin\ \varepsilon + \sin\ i\ \cos\ \varepsilon
\end{aligned} \tag{3.29}$$

where Ω and i are the orbital elements, as defined earlier, ε is the *obliquity of the ecliptic*, and the quantities α', β', γ' (which Gauss called sin a, sin b, and sin c, respectively; all must have positive, non-zero values), and A, B, and C, need to be computed only once per orbit. From these equations, the heliocentric equatorial coordinates become

$$\begin{aligned}
x &= r\alpha'\ \sin\ (A + \omega + \upsilon) \\
y &= r\beta'\ \sin\ (B + \omega + \upsilon) \\
z &= r\gamma'\ \sin\ (C + \omega + \upsilon)
\end{aligned} \tag{3.30}$$

where ω is the argument of perihelion. Finally, the geocentric equatorial coordinates may be computed:

$$\begin{aligned}\xi &= x + X\\ \eta &= y + Y\\ \zeta &= z + Z\end{aligned} \tag{3.31}$$

where X, Y, and Z are the rectangular coordinates of the Sun in geocentric equatorial coordinates, tabulated annually in the Astronomical Almanac as quantities labeled "x," "y," and "z," respectively. The relationship between these quantities and right ascension and declination is given by

$$\begin{aligned}&\rho \cos\delta\ \cos\alpha = \xi\\ &\rho \cos\delta\ \sin\alpha = \eta\\ &\rho \sin\delta = \zeta\end{aligned} \tag{3.32}$$

where ρ is the geocentric distance of the planet:

$$\rho = \left[\xi^2 + \eta^2 + \zeta^2\right]^{1/2} \tag{3.33}$$

so,

$$\sin\delta = \zeta/\rho \quad \text{and} \quad \tan\alpha = \eta/\xi \tag{3.34}$$

More detailed treatment can be found in Moulton (1914/1958, pp. 182–189) or in other texts on Celestial Mechanics [e.g., Brouwer and Clemence (1961), Danby (1962/1964, 1988), Neutsch and Scherer (1992), Smart (1953), or Szebehely (1989)], and, in somewhat more abridged but focused discussions, in Montenbruck (1989) or Schlosser et al. (1991/4).

3.5 Kepler's Equation[5]

A useful starting point is the *vis-viva* equation, which may be derived from the energy equation (3.15). At perihelion, the energy equation may be written as

$$E = (1/2)\left(\dot{r}^2 + r^2\dot{\theta}^2\right) - (\mu/r) = (1/2)\left(0 + r^2\dot{\theta}^2\right) - (\mu/r) \tag{3.35}$$

and because $r^2\dot{\theta}^2 = h^2/r^2$, recalling that $h^2/\mu = a(1 - e^2)$, and that at perihelion, $r = a(1 - \mathrm{e})$ [see (3.24) and the text near it], we can reduce (3.35) to

[5] Note that *Kepler's equation* should not be confused with any of Kepler's three laws.

$$E = -(1/2)\mu/a \tag{3.36}$$

E is a constant, so it has the same value everywhere in the orbit. Substituting and setting v^2 equal to the sum of the squared velocity terms in (3.15), we get the *vis-viva* equation:

$$v^2 = \mu[(2/r) - (1/a)] \tag{3.37}$$

If we now make use of Kepler's third law, sometimes referred to as "Kepler III,"

$$a^3 = \left[\mu/\left(4\pi^2\right)\right]P^2 \tag{3.38}$$

and define the *mean* (angular) *motion*, n, by

$$n \equiv 2\pi/P \tag{3.39}$$

the latter becomes

$$n = \mu^{1/2}/a^{3/2} \tag{3.40}$$

Because, in general,

$$\begin{gathered} v^2 = \dot{r}^2 + \left(r\dot{\theta}\right)^2 \\ \left(r\dot{\theta}\right)^2 = h^2/r^2 \end{gathered}$$

and

$$h^2/\mu = a\left(1 - e^2\right)$$

we have,

$$v^2 = \dot{r}^2 + h^2/r^2 = \dot{r}^2 + \mu\left(a/r^2\right)\left(1 - e^2\right) \tag{3.41}$$

Also, by squaring (3.40), and rearranging, we get

$$\mu = n^2a^3 \tag{3.42}$$

and, with this substitution into (3.37),

$$v^2 = n^2a^3(2a - r)/(ar) \tag{3.43}$$

Then, solving for the square of the derivative of the distance scalar on the right-hand side of (3.41), and substituting for μ from (3.42) and for v^2 from (3.43), we obtain

$$\dot{r}^2 = \left[n^2 a^2 (2a - r) r - n^2 a^4 \left(1 - e^2\right)\right] / r^2 \tag{3.44}$$

From this,

$$\begin{aligned} n\,dt &= (r/a)[2ar - r^2 - a^2(1 - e^2)]^{-1/2} dr \\ &= (r/a)\left[a^2 e^2 - (a - r)^2\right]^{-1/2} dr \end{aligned} \tag{3.45}$$

We now define the *eccentric anomaly*, traditionally written as E (but not to be confused with the orbital energy), such that

$$r = a\,(1 - e \cos E) \tag{3.46}$$

From this, (a − r), r/a, and the differential d$r = ae$ sin E dE are obtained. After substitution in (3.45),

$$n\,dt = (1 - e \cos E)\,dE \tag{3.47}$$

which, when integrated, yields

$$n\,(t - T_0) = E - e \sin E \tag{3.48}$$

where T_0 is the initial value of time (an instant when $E = 0$, i.e., at pericenter), and t the terminal time, the instant of interest.

Then the *mean anomaly*, M [defined in (3.26)], may be written as

$$M = n\,(t - T_0) \tag{3.49}$$

With (3.48), *Kepler's equation* emerges:

$$M = E - e \sin E \tag{3.50}$$

Figure 3.4 illustrates how the eccentric anomaly, E, is related to the true anomaly, υ. Note that neither E nor υ increases at a linear rate as the orbiting object moves in its orbit. Like E, M is measured from the center, but it increases at a constant rate. The geometry of this figure may be used to derive Kepler's equation geometrically, and we now proceed to do so.

The geometrical interpretation of Kepler's equation may be demonstrated as follows.

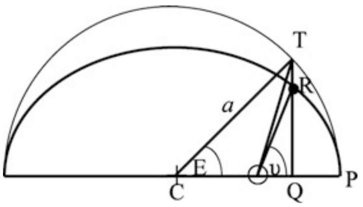


Fig. 3.4 The relationship between υ and E. Note that TP is a circular arc; RP is an elliptical arc

Because M is an angle that increases uniformly with time, and because the areal velocity is constant [refer to the text near equations (3.11) and (3.12)],

$$M/(2\pi) = \text{area P} \odot \text{R/ellipse area} = \text{area T} \odot \text{P/circle area}$$

but

$$\text{area T} \odot \text{P} = \text{area TCP} - \text{area TC}\odot$$

The area of the segment of the circle is

$$\text{area TCP} = (1/2)a(aE)$$

where the angular quantity E, when treated algebraically outside of trigonometric functions, as here, must be expressed in radian measure. The line segment $\text{C}\odot = ae$, so the triangular area

$$\text{TC}\odot = (1/2)(ae)\text{TQ} = (1/2)(ae)a \sin E$$

Therefore,

$$\text{area T} \odot \text{P} = (E/2)a^2 - (1/2)(ae)a \sin E$$

whence,

$$M/(2\pi) = \left(a^2/2\right)(E - e \sin E)/\left(\pi a^2\right)$$

or

$$M = E - e \sin E$$

which is Kepler's equation, once again. The study of this equation has a long and venerable history. See Neutsch and Scherer (1992), especially Chaps. 3 and 4, for both history and method.

It is necessary to solve Kepler's equation for E in order to obtain the position of the object at the instant t. The relation between υ and E is obtained by setting equal equations (3.28) and (3.46):

$$a\left(1-e^2\right)/(1+e\,\cos\upsilon)=a(1-e\,\cos E)$$

whence,

$$\cos\upsilon=(\cos E-e)/(1-e\,\cos E) \tag{3.51}$$

Unlike (3.27), (3.51) yields the true anomaly regardless of the size of the eccentricity.

There are several ways to achieve solutions for Kepler's equation, some more easily, others more precisely. Some of these are:

1. Graphically, one can plot the functions $f = \sin E$ and $g = (1/e)(E - M)$. The intersection of the two curves is the solution.
2. Tabularly, one can examine, perhaps interpolate, among these quantities to find the solution.
3. Iteratively, which is best suited for computation. We illustrate:

   (a) Start with a rough solution, say E_0.

   (b) Compute

$$M_0=E_0-e\sin E_0 \tag{3.52a}$$

   (c) But M is known, so compute the difference

$$\Delta M=M-M_0 \tag{3.52b}$$

   (d) Then, differencing (3.50),

$$\Delta M=\Delta E-(e\,\cos E)\Delta E \tag{3.52c}$$

   compute the correction,

$$\Delta E_0=\Delta M/(1-e\,\cos E) \tag{3.52d}$$

   (e) Then find a new value for E:

$$E_1=E_0+\Delta E_0 \tag{3.52e}$$

   (f) Increase the subscripts by one, and repeat steps (b), through (e); then continue iterating until ΔM matches the precision in M.

Examples can be found in Danby (1962, p. 148ff, 1988, p. 149ff) and in Moulton (1914/1958, p. 160ff, 181 (#3)).

It may be shown further that E is expressible as an expansion of M, resulting in a series. The expression recapitulates the iterative procedure to a given order.

3.6 Uses and Limitations of Two-, Three- and *n*-Body Solutions

First we illustrate a use for the two-body system, one that has generated billions of dollars of profit for communications satellite corporations.

Example 3.1 Use of two- and restricted three-body system celestial mechanics.

The two-body equations of motion suffice for approximate solutions in the presence of a dominating mass and in the absence of major sources of perturbations. For example, consider the case of a synchronous satellite. It is possible to launch a small satellite with a rocket into low Earth orbit and then to transfer it to another, higher orbit. Higher orbits have the advantage of little atmospheric drag on the satellite, which causes a loss of orbital energy, and eventual orbital decay. A particular type of high orbit has another great advantage for communications: a synchronous satellite (at present found only around the Earth, where it is called a *geo-synchronous satellite*) orbit, illustrated in Fig. 3.5.

We have already mentioned that two-body systems, although analytically rigorous to describe, are not sufficient to describe the mechanics of objects moving in the solar system to the highest precision, in most cases.

The *restricted three-body* problem has a closed or analytical solution. This is the only exact solution for the interaction of three bodies and it requires special circumstances: one of the three bodies must have such a sufficiently small mass that it has no significant effect on the other two bodies.

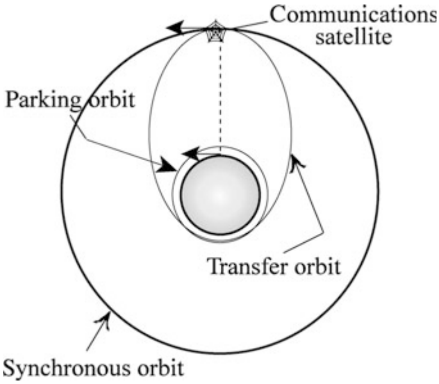


Fig. 3.5 The synchronous satellite and transfer orbits. An object in a geosynchronous orbit has the same period, and thus angular speed, as an object on the Earth's surface, subjected to only the Earth's rotation

We mentioned in Sect. 3.3 the solutions for the restricted three-body problem and refer to them again later, but it is not the only possible way to solve multi-body interactions. Numerical calculations for *n* bodies can be carried out and, in this way, orbits can be plotted and object positions predicted, for the major bodies of the solar system. This is how space mission trajectories are computed, for example, even in the presence of major perturbation sources.

Example 3.2 Use of *n*-body calculations.
The basic idea for *n*-body orbital calculations is this:

1. The distances of the object in question from all significant masses are established for some instant, from the osculating orbit of the object perhaps, and the theories of motion of all the other bodies.
2. Then the force due to each of the objects is computed to find the total force, and the net acceleration of the body found.
3. The acceleration (multiplied by a small time step) is then applied as a vector addition to the previous velocity of the body.
4. The mean velocity over the time step (a small interval) is calculated and the net change in position computed; return to stage (1).

See Brouwer and Clemence (1961, p. 171ff) for an example, the procedure for computing the ephemeris of the asteroid *1 Ceres*.

Modern computers have greatly aided the process of integration that is involved in predicting positions accurately. As with most intensive and repeated operations, the limiting precision at each step determines the accuracy of the end result.

A number of long-term integrations of planetary orbits have been undertaken since the JPL (Jet Propulsion Lab.) ephemeris DE102 computed a consistent set of planetary positions over the interval 1411 B.C. to 3002 A.D. (Newhall et al. 1983); Subsequently, integrations were carried out for intervals of 100 My (LONGSTOP by Nobili 1988), 200 My (MIT project by Applegate et al. 1986), 845 My (Digital Orrery Project by Sussman and Wisdom 1988), 1 Gy (by Wisdom and Holman 1991), and 5 Gy (Laskar and Robutel 2001; and, with higher order effects, by Laskar and Gastineau 2009). These have been used to study the stability of the planetary orbits and their resonances.

A major difficulty in multi-body orbital predictions is the presence of families of orbits where chaos can cause unpredictability. A circumstance can arise wherein a very small change in the force results in a significantly different motion. Some asteroid, cometary, and satellite orbits, and even those of dwarf planets (e.g., Pluto) are subject to this. Because of this, in some cases we cannot really be sure what the remote past was like, and we can be even less secure about the future. The past at least has the constraint that the endpoint of the orbital evolution is the present configuration; in the worst cases, we may not know what the configuration will be in the distant future.

Now, however, we consider an interesting problem: the transfer of angular momentum from orbital motion into rotational motion, and vice versa.

3.7 Spin–Orbit Coupling

The coupling between rotation and orbit requires some interaction other than the inverse square law operating between two effective mass centers. The most common such interaction is tidal.

3.7.1 Effect of Tidal Friction on Rotation

The tidal interaction between two objects with center-to-center distance r can be expressed as

$$\begin{aligned}\Delta F &= -G\mathfrak{M}m\left[(r-\Delta r)^{-2} - r^{-2}\right] \\ &= -G\mathfrak{M}m\left[2r\Delta r - (\Delta r)^2\right]r^{-2}\left[r^2 - 2r\Delta r + (\Delta r)^2\right]^{-1}\end{aligned} \tag{3.53}$$

where ΔF is the difference between the gravitational attraction at the center and the closer edge of one of the objects by the other. These two points are a distance Δr apart, equal to the radius of the object subject to the tides. If we can assume, say, that $\Delta r \lll r$, we need retain only first order terms; dividing by r^2 we then get

$$\begin{aligned}\Delta F &\approx -G\mathfrak{M}m[(2\Delta r/r)][r^2 - 2r\Delta r]^{-1} \\ &\approx 2G\mathfrak{M}m\Delta r/r^3\end{aligned} \tag{3.54}$$

The result can be obtained and expressed more elegantly by use of the *del* or *nabla* operator as

$$\Delta F = \nabla\left(-G\mathfrak{M}mr^{-2}\right) = 2G\mathfrak{M}m\Delta r/r^3 \tag{3.55}$$

If we are examining the bulge raised in the Earth by a satellite, $\Delta r = R_\oplus$. Because the distance of the Moon varies by $\pm ae = \pm 0.055a$, the relative change in the tide-raising force across the orbit is

$$\delta(\Delta F)/\Delta F \approx 0.165$$

Example 3.3 Relative tides due to the Sun and Moon on the Earth.
The tidal effect on the Earth by the Moon is currently about twice that by the Sun:

$$\Delta F_{\oplus-\odot} = -2Gm_\oplus\mathfrak{M}_\odot R_\oplus/a_\oplus^3 \tag{3.56}$$

and

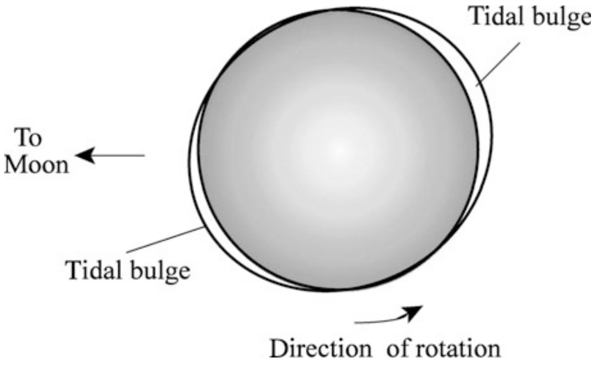


Fig. 3.6 Tidal braking of the Earth by lunar torques

$$\Delta F_{\oplus-☽} = -2Gm_{\oplus}m_{☽}R_{\oplus}/a_{☽}^{3} \tag{3.57}$$

where m and $\mathfrak{M}$ are masses, R is a radius, and a is a semi-major axis; the symbols of Sun, Moon, and Earth indicate which is which. Then, dividing (3.57) by (3.56),

$$\Delta F_{\oplus-☽}/\Delta F_{\oplus-\odot} = [m_{☽}/\mathfrak{M}_{\odot}](a_{\oplus}/a_{☽})^{3}$$
$$= \left[7.35 \times 10^{22}/2.0 \times 10^{30}\right]\left(1.5 \times 10^{11}/3.8 \times 10^{8}\right)^{3} = 2.2$$

Notice that there are bulges raised on both sides of the planet by tidal action. The far side is accelerated less than, say, the planetary core, while the core is accelerated less than the near side.

Because of friction, however, the tidal bulges may not be able to localize to the sub-satellite position on the planet. If the angular speed of rotation exceeds the angular motion of the satellite in its orbit, and is in the same direction, then friction may result in the tide being swept along with the rotation, causing the bulge to precede the satellite. This is the case for the Earth (see Fig. 3.6, where the tides have been greatly exaggerated for visibility). The net effect is for the satellite to drag on the near-side bulge and to accelerate the far-side bulge. The slightly greater proximity of the near-side bulge (by the inverse square law) results in the torque on the near side being larger than that on the far side. The differential force again goes as $r^{-\ 3}$. This causes a net braking of the rotation of the planet. The evidence for the slowing of the Earth's rotation is overwhelming, as we note in the next section. The heat generated in the Earth by tidal friction is discussed in Sect. 6.1.1.2.

3.7.2 Effect of Tidal Friction on Orbits

The same bulge that gets braked in the discussion in the preceding section also accelerates the Moon. The acceleration increases the instantaneous orbital speed, resulting in a slight increase in the semi-major axis. As the orbit enlarges the orbital speed decreases. Thus the second consequence of tidal friction is to increase the orbit of the Moon, and thus, by Kepler's third law, its period.

The evidence for secular variation in the semi-major axis of the Moon is also strong. Consolmagno and Schaefer (1994, p. 246) cite evidence that in the Devonian Period [~400 million years (My) before present] there were only 10 days in the lunar month, with an average change in the length of the day since then of 25^{s}/My. They estimate that the Moon at that time would have been half as distant as it is now. The rate of recession depends strongly on the effectiveness of tidal friction, which in the current epoch is concentrated in shallow sea passages such as the Bering Strait and the English Channel. This means that the rate must vary over time because shallow seas come and go over geological time due to continental drift.

Therefore, continental drift has changed the rate of spindown of the Earth and the rate of recession of the Moon. Coupled to this must be the mean ocean level, which, in turn, depends on the state of glaciation and ice cap thickness.

The end of the current trend (Earth rotation slowing; Moon receding) will occur when the Earth and Moon become tidally locked, with the Earth day and lunar month being of approximately equal length. However, this equilibrium situation may not last. The tidal action of the Sun is to slow the Earth's rotation also, and that will continue, presumably. Once the Earth-Moon lock becomes broken, the slower Earth may brake the Moon's motion, resulting in an inwardly spiraling Moon. Such a decay can be seen in Phobos' orbit about Mars.

Differential forces have importance beyond the tides raised on solid bodies. They can, for instance, result in disruption of a planetary satellite's orbit or even destruction of the satellite; by extension, the "satellite" could be a planet acted on by another star, or a star cluster acted on by another galaxy. We take up the first of these interesting cases in Chap. 13 and the second in Chap. 16 of Milone and Wilson (2014).

3.7.3 Resonances and Commensurabilities

Tide-raising effectiveness must change with time as the distance increases. When the rotation and revolution are in some type of synchronism, a spin-orbit resonance is said to occur. An example is seen in the 1:1 ratio of revolution to rotation periods of Earth's Moon. However, the ratio need not be 1. For Mercury, the ratio is 3:2 (see Sect. 9.1.3).

There are also orbit-orbit resonances. The classical example of the Galilean satellites was discovered by Pierre-Simon de Laplace (1749–1827) in 1805. In this case, $P_{\mathrm{Ganymede}} = 7^{\mathrm{d}}.1455$; $P_{\mathrm{Europa}} = 3^{\mathrm{d}}.55118$, and $P_{\mathrm{Io}} = 1^{\mathrm{d}}.76914$, hence

$P_{\mathrm{Ganymede}} = 2P_{\mathrm{Europa}}$, and $P_{\mathrm{Europa}} = 2P_{\mathrm{Io}}$. Expressing the mean motion as $n = 2\pi/P$, the mutual locking can be described by the relation

$$n_{\mathrm{Io}} - 3n_{\mathrm{Europa}} + 2n_{\mathrm{Ganymede}} = 0 \tag{3.58}$$

Close approximations to Laplace resonances are now being seen in extrasolar planets. In the multi-planet system GJ 876 (see Chap. 1; and Table 16.2 in Milone and Wilson 2014), the periods of the three outer planets are: 30.09d, 61.12d, and 124.26d.

Table 3.1 Some solar system spin-orbit resonances

Objects	e	i	P_{rtn}	P_{rev} (Sid.)	P_{rev} (Syn)	P_1/P_2 = ?
☿-⊕	0.206	$7\overset{\circ}{.}003$	59d	$87\overset{\mathrm{d}}{.}969$	$115\overset{\mathrm{d}}{.}88$	0.509 ≈ 1/2(rtn/syn)
☿-☉						0.671 ≈ 2/3(rtn/sid)
♀-⊕	0.007	$3\overset{\circ}{.}395$	$244\overset{\mathrm{d}}{.}3$	$224\overset{\mathrm{d}}{.}701$	$583\overset{\mathrm{d}}{.}92$	0.418 Earth ≈ 5/12 (rtn/syn)

Table 3.2 Some solar system orbit-orbit resonances

Objects	e	i	P_{rev}	P_1/P_2
Earth	0.017	$0\overset{\circ}{.}0$	$1\overset{\mathrm{y}}{.}0$	0.625 ≈ 5/8
Toro	0.435	$9\overset{\circ}{.}3$	1.6	
Earth	0.017	$0\overset{\circ}{.}0$	1.0	0.393 ≈ 2/5
Ivar	0.397	$8\overset{\circ}{.}3$	2.545	
Jupiter	0.048	$1\overset{\circ}{.}4$	11.86	1.501 ≈ 3/2
Hilda	0.15	$7\overset{\circ}{.}9$	7.90	
Jupiter	0.048	$1\overset{\circ}{.}4$	11.86	1.333 ≈ 4/3
Thule	0.03	$23\overset{\circ}{.}$	8.90	
Jupiter	0.048	$1\overset{\circ}{.}4$	11.86	1
Trojans	0.15:	15:	11.86	
Tethys	0.00	$1\overset{\circ}{.}1$	$1\overset{\mathrm{d}}{.}887802$	2.003
Mimas	0.020	$1\overset{\circ}{.}5$	0.942422	
Dione	0.002	$0\overset{\circ}{.}0$	$2\overset{\mathrm{d}}{.}93681$	1.997
Enceladus	0.0045	$0\overset{\circ}{.}0$	1.37028	
Hyperion	0.104	$0\overset{\circ}{.}5$	$21\overset{\mathrm{d}}{.}27666$	1.334 ≈ 4/3
Titan	0.0290	$0\overset{\circ}{.}3$	15.945452	
Pluto	0.247	$17\overset{\circ}{.}1$	$248\overset{\mathrm{d}}{.}43$	1.508 ≈ 3/2
Neptune	0.0087	$1\overset{\circ}{.}5$	164.78	

Finally, there is a host of resonances among Saturn's moons. Tables 3.1 and 3.2 list the best known spin-orbit and orbit-orbit coupling cases in the solar system.

Among longer period effects is the *long-period inequality* between Jupiter and Saturn; their motions are not quite commensurable, but very close to it:

$$5n_{\mathrm{Saturn}} - 2n_{\mathrm{Jupiter}} = 3\overset{\prime\prime}{.}99/\mathrm{d} \tag{3.59}$$

where $n_{\mathrm{Jupiter}} = 299\overset{\prime\prime}{.}13/\mathrm{d}$ and $n_{\mathrm{Saturn}} = 120\overset{\prime\prime}{.}45/\mathrm{d}$. Because this inequality involves mean motions, and in one revolution, $n = 1{,}296{,}000''/\mathrm{P(y)}$, one may also write it in terms of the periods of revolution in years,

$$5P_{\mathrm{J}} - 2P_{\mathrm{S}} \overset{\circ}{=} \kappa\, P_{\mathrm{J}}P_{\mathrm{S}} \tag{3.60}$$

where $P_{\mathrm{J}} = 11\overset{\mathrm{y}}{.}862$ and $P_{\mathrm{S}} = 29\overset{\mathrm{y}}{.}459$, and $\kappa = 0.00112$. The terms on the LHS have the values $58\overset{\mathrm{y}}{.}918$ and $59\overset{\mathrm{y}}{.}310$, respectively. Thus, Jupiter and Saturn will come to the same heliocentric configuration about every 59 years.

In the Saturn system there are a number of commensurabilities. For instance, $n_{\text{titan}} = 1.334342\ n_{\text{Hyperion}}$, very close to a 4:3 resonance.

In some cases the perturbations on small objects are sufficient to lock them in or sometimes out of certain orbits. Dione and Tethys have a 1:1 resonance with small objects in their respective orbits, and Janus and Epimetheus have horseshoe orbits and interchange orbits, and are also locked into a 1:1 orbital resonance. The moons of Saturn sweep away material from certain ring tori, leaving gaps. In the Uranian system, Rosalind and Cordelia have a 5:3 resonance. Cordelia and Ophelia provide bounds to the ε ring, with 24:25 and 14:13 resonances. Pluto and Charon are also locked into a 1:1 resonance.

Jupiter has a 1:1 orbital resonance with the two sets of "Trojan" asteroids located ~60° fore and aft of it. Jupiter also perturbs the orbits of minor planets with sub-multiples of Jupiter's period. The "Kirkwood gaps" in the asteroid belt are the result (See Milone and Wilson 2014 Ch. 15.7.2).

Spin-orbit and orbit-orbit resonances are relatively short period: the 3:2 resonance between Pluto's 249-year orbit and Neptune's 166-year orbit is the longest currently known in our solar system. Other, much longer-period, *secular resonances* can also be important; e.g., the inner edge of the main asteroid belt is determined by the ν_6 ("nu-6") secular resonance, of period ~50,000 years, in which the rate of precession of an asteroid's longitude of perihelion equals the rate of precession of Saturn's longitude of perihelion. This secular resonance is discussed again in Milone and Wilson (2014), Sect. 15.7.5. Finally, resonances seem to be common among extrasolar planets as well (Milone and Wilson 2014, Chap. 16). In some systems, numerical simulations seem to suggest that dynamical stability over intervals of millions of years is associated with planets having commensurable orbits.

3.8 Perturbations

3.8.1 Causes and Effects

Sources of perturbations include other bodies, non-spherical distributions of the mass of the objects in the two-body system or both, and viscous media or other sources of drag. The orbital elements of an object in a two-body orbit subject to a point source mass will remain constant. However, as noted in Sect. 3.2, subject to perturbations they will undergo variation. The variation of each element depends on the relative direction and magnitude of the perturbing force. At some instant, the object in such a perturbed orbit may be said to have an *osculating orbit* because it "kisses" the true path at that instant (and perhaps at none other). If all perturbing forces were to disappear at that instant (and remained absent thereafter), the orbital motion thenceforth would be described accurately by the elements of the osculating orbit.

Relative to the two-body orbital plane, perturbations that are said to be

1. *Normal*, alter the inclination (i) (and, through precession, Ω and ω);
2. *Radial*, alter the eccentricity (e), and possibly the semi-major axis (a); and
3. *Transverse* to the other two directions, alter a and e.

Transitions to/from transfer orbits, when a thrust is given the spacecraft at either pericentre or apocentre, involve perturbing effect 3.

Satellite orbits undergoing perturbations are subject to effects 1 and 2, and precess.

We now determine the changes in the osculating elements due to a general perturbation force, but which we assume to be small compared to the principal two-body force. This is usually not an unrealistic assumption, given the relatively large separations of bodies and the dominance of the Sun's gravitational effects in the solar system. An excellent detailed source for this treatment is Murray and Dermott (1999/2001), which our summary follows.

3.8.2 Deriving the Variation in the Orbital Elements

A small disturbing force per unit mass, dF, sometimes referred to as the *perturbative acceleration*, may be expressed in terms of its components:

$$\mathrm{d}\mathbf{F} = R\hat{\mathbf{r}} + T\hat{\boldsymbol{\theta}} + N\hat{\mathbf{z}} \tag{3.61}$$

where $\hat{\mathbf{r}}$ and $\hat{\boldsymbol{\theta}}$ are unit vectors as defined in Sect. 3.1 and $\hat{\mathbf{z}}$ is in the direction normal to the orbit; and R, T, and N are the components of the perturbative force in the radial, transverse, and normal directions, respectively.

When perturbations occur, the energy of the orbit is no longer constant. The time-variation is

$$\dot{E} = \dot{\mathbf{r}} \cdot \mathrm{d}\mathbf{F} = \dot{r}R + r\dot{\theta}T \tag{3.62}$$

As per (3.36), $E = -(1/2)\mu/a$, so the variation in a is

$$\dot{a} = 2a^2\dot{E}/\mu \tag{3.63}$$

Rewriting (3.25) for the ellipse explicitly, we get

$$r = a\left(1 - e^2\right)/[1 + e\cos\theta] \tag{3.64}$$

Noting that the orbit angle variation, θ, is that of the true anomaly, υ, setting $\dot{\theta} = \dot{\upsilon}$, and taking the time derivative due to motion in the orbit (a and e kept constant),

$$\dot{r} = r\dot{\upsilon}\, e \sin \upsilon / [1 + e \cos \upsilon] \tag{3.65}$$

and with $(r\dot{\upsilon}) = h^2/r^2$,
and $h^2 = a\mu(1 - e^2) = n^2a^4(1 - e^2)$,
where $n = 2\pi/P = [\mu/a^3]^{½}$ [from (3.39), (3.40)], we can show that

$$\dot{r} = n\, a\, e \sin \upsilon / \left(1 - e^2\right)^{1/2} \tag{3.66}$$

and

$$r\dot{\upsilon} = n\, a\, (1 + e \cos \upsilon) / \left(1 - e^2\right)^{1/2} \tag{3.67}$$

Substituting (3.66) and (3.67) into (3.62), from (3.63) we derive

$$\dot{a} = \left\{2/\left[n\sqrt{1 - e^2}\right]\right\}[(e \sin \upsilon)R + (1 + e \cos \upsilon)T]$$

or, with a explicit,

$$\rightarrow \qquad \dot{a} = \left\{2\, a^{3/2}/\sqrt{\mu\,(1 - e^2)}\right\}[(e \sin \upsilon)R + (1 + e \cos \upsilon)T] \tag{3.68}$$

From the relationship $h^2/\mu = a(1 - e^2)$, we get

$$e = \left\{1 - h^2/(\mu\, a)\right\}^{1/2} \tag{3.69}$$

whence,

$$\dot{e} = (1/2)\ \left\{1 - h^2/(\mu\, a)\right\}^{-1/2}\left[-2h\dot{h}/(\mu\, a) + h^2\,\dot{a}/\left(\mu\, a^2\right)\right] \tag{3.70}$$

We know $\dot{a}$ from (3.68) but we must find $\dot{\mathbf{h}}$. From (3.11a),

$$\mathbf{h} = \mathbf{r} \times \dot{\mathbf{r}}$$

It follows that

$$\dot{\mathbf{h}} = \dot{\mathbf{r}} \times \dot{\mathbf{r}} + \mathbf{r} \times \ddot{\mathbf{r}} = 0 + \mathbf{r} \times \mathrm{d}\mathbf{F} \tag{3.71}$$

and because

$$\mathbf{r} = r\,\hat{\mathbf{r}} \tag{3.72}$$

with (3.61),

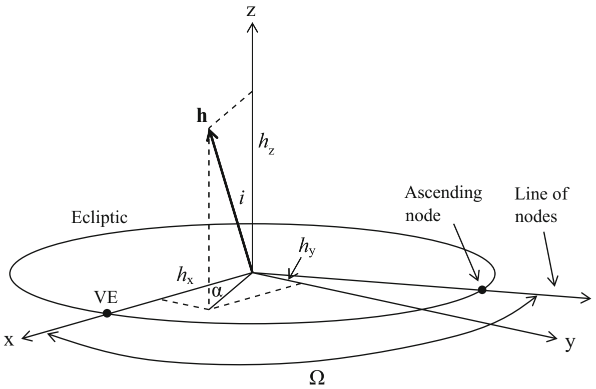


Fig. 3.7 The components of **h** in the x, y and z directions. VE = vernal equinox; Ω = longitude of ascending node; i = inclination of the orbit to the reference plane (here, the ecliptic)

$$\dot{\mathbf{h}} = r\hat{\mathbf{r}} \times \left(R\hat{\mathbf{r}} + T\hat{\boldsymbol{\theta}} + N\hat{\mathbf{z}}\right) \tag{3.73}$$

so that

$$\dot{\mathbf{h}} = rT\hat{\mathbf{z}} - rN\hat{\boldsymbol{\theta}} \tag{3.74}$$

However, the $r\,N\,\hat{\theta}$ term alters the direction but not the magnitude of **h**, hence

$$\dot{h} = rT \tag{3.75}$$

Substituting (3.75), (3.68), (3.46) and the relation, $h^2 = a\mu(1 - e^2)$, from (3.24), into (3.70), after some simplification, we arrive at

$$\rightarrow \qquad \dot{e} = \left[a\left(1 - e^2\right)/\mu\right]^{1/2}[R\sin\upsilon + T(\cos E + \cos\upsilon)] \tag{3.76}$$

where this E is the eccentric anomaly. Note that the variation of the eccentricity depends only on the components of the perturbation in the orbital plane.

The variation of the inclination, i, is more complicated. We begin with the components of **h** in the orthogonal directions $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, $\hat{\mathbf{z}}$ in which $\hat{\mathbf{x}}$ is in the direction of the vernal equinox, $\hat{\mathbf{y}}$ is 90° in the direction of increasing longitudinal coordinate, in the reference plane, and $\hat{\mathbf{z}}$ is normal to the plane. The components of the angular momentum vector in the reference frame are the dot products of **h** and the unit vectors of the x, y, and z directions: $h_z = \mathbf{h} \bullet \hat{\mathbf{z}}$, etc.

Figure 3.7 illustrates the geometry. To simplify the figure, the orbital plane of the body is not shown, but it is inclined at angle i to the reference plane (the ecliptic in Fig. 3.7), intersects the reference plane along the line of nodes, and is perpendicular to **h**. The longitude of the ascending node is Ω, measured from the x-axis (the vernal equinox, VE). When two planes intersect at an angle, their perpendiculars intersect at the same angle, so **h** is inclined by angle i to the z-axis. Figure 3.7 also shows the projection of **h** onto the x, y plane, $h_{xy} = h \sin i$. The dashed line joining **h** to the z-axis (equal to the line h_{xy}) is perpendicular to the line of nodes, so, labeling the angle between h_{xy} and the x-axis α, we have $\alpha = \Omega - 90°$. Then

$$\begin{aligned} h_x &= h_{xy} \cos\alpha = h \sin i \cos(\Omega - 90°) = h \sin i \sin\Omega \\ h_y &= h_{xy} \sin\alpha = h \sin i \sin(\Omega - 90°) = -h \sin i \cos\Omega \\ h_z &= h \cos i \end{aligned} \tag{3.77}$$

Taking the derivative of h_z,

$$\dot{h}_z = \dot{h} \cos i - h \sin i \ \ i' \tag{3.78}$$

where $i' = di/dt$, so

$$i' = [\dot{h} \cos i - \dot{h}_z]/(h \sin i)$$

Simplifying,

$$\begin{aligned} i' &= [(\dot{h}/h) \cos i - \dot{h}_z \cos i/h_z]/\sin i \\ i' &= [(\dot{h}/h) - (\dot{h}_z/h_z)]/\sqrt{(1 - \cos^2 i)/\cos^2 i} \end{aligned}$$

From (3.77), we get

$$i' = [(\dot{h}/h) - (\dot{h}_z/h_z)]/\sqrt{(h/h_z)^2 - 1} \tag{3.79}$$

In terms of the disturbing force components given in (3.61), the derivatives of h_x, h_y, and h_z may be shown to be

$$\begin{aligned} \dot{h}_x &= r[T \sin\ i \sin\ \Omega + N \sin(\omega + \upsilon) \cos\ \Omega + N \cos\ (\omega + \upsilon) \cos\ i \sin\ \Omega] \\ \dot{h}_y &= \mathrm{r}[-T \sin\ i \cos\ \Omega + N \sin\ (\omega + \upsilon)\ \sin\ \Omega - N \cos\ (\omega + \upsilon) \cos\ i \cos\ \Omega] \\ \dot{h}_z &= r[T\ \ \cos\ i - N\ \ \cos\ (\omega + \upsilon) \sin\ i] \end{aligned} \tag{3.80}$$

which involve transformations from the orbital plane into the reference plane.

Note that the argument ($\omega + \upsilon$) is the "argument of latitude," u, mentioned in Sect. 3.4, and it is measured from the node. Ultimately, we get

$$i' = rN \cos(\omega + \upsilon)/h$$

or

$$\rightarrow \qquad i' = \left\{[a(1 - e^2)/\mu]^{1/2} N \cos(\omega + \upsilon)\right\}/[1 + e \cos\upsilon] \tag{3.81}$$

Note that only the component of the perturbing force normal to the orbital plane has an effect on i.

The effect on the longitude of the ascending node is obtained by dividing the first by the second equation of (3.77)

$$h_x/h_y = -\tan \ \Omega \tag{3.82}$$

then, taking the derivative,

$$-\sec^2\Omega \cdot \dot{\Omega} = \left(\dot{h}_x/h_y\right) - \left(h_x\dot{h}_y/h_y^2\right)$$

after manipulation,

$$\dot{\Omega} = \left(h_x\dot{h}_y - h_y\dot{h}_x\right)/\left(h^2 - h_z^2\right) \tag{3.83}$$

and, after appropriate substitutions,

$$\dot{\Omega} = rN\sin\left(\omega + \upsilon\right)/\left(h\sin i\right)$$

or

$$\rightarrow \qquad \dot{\Omega} = \left\{\left[a\left(1 - e^2\right)/\mu\right]^{1/2} N\sin\left(\omega + \upsilon\right)\right\}/[\sin i(1 + e\cos\upsilon)] \tag{3.84}$$

Note the sensitivity of $\dot{\Omega}$ only to the N component of d**F**. This component causes the motion of the plane of the orbit.

We can find the variation of ω by the substitution $\upsilon = u - \omega$ in the equation of the ellipse, where u is the argument of the latitude, referred to above and in Sect. 3.4. The expression takes on the form

$$h^2 = \mu r\left\{1 + \sqrt{1 + 2Eh^2/\mu^2}\ \cos\left(u - \omega\right)\right\} \tag{3.85}$$

Now we need to take the time derivative. For this exercise, we are interested in the instantaneous value of $\dot{\omega}$ due to the perturbing force, so r is treated as a constant even though E, h, and ω are allowed to vary. The result is

$$\begin{aligned}\dot{\omega} = 2h\dot{h}\left[(1/r) + E\cos\left(u - \omega\right)/\left(e\mu\right)\right]/\left[e\mu\sin\left(u - \omega\right)\right] + \dot{u}\\ - [h/(e\mu)]^2\dot{E}\cot\left(u - \omega\right)\end{aligned} \tag{3.86}$$

Making use of substitutions for $\dot{h}$, $\dot{E}$ [and then $\dot{a}$ from (3.63)] as given above, we get

→
$$\begin{aligned}\dot{\omega} &= (1/e)[a(1-e^2)/\mu]^{1/2}\{-R\cos\upsilon \\ &\quad + T\sin\upsilon\left[(2+e\cos\upsilon)/\left(1+e\cos\upsilon\right)\right]\} - \dot{\Omega}\cos i\end{aligned} \tag{3.87}$$

The last term, involving $\dot{\Omega}$, arises from the changing position of the nodes, from which u is measured.

The effect of a perturbation on the epoch (the instant that the object is at a particular point in the orbit, such as the pericenter), can be determined by differentiation of Kepler's equation, defined in (3.46), and rewritten here as

$$M = \varepsilon - e\sin\varepsilon \tag{3.88}$$

where ε is the eccentric anomaly as we defined it in (3.46), but to avoid confusion with the energy per unit mass in this section, we make the substitution $\varepsilon = E$ for the eccentric anomaly. Now $M = n(t - \mathrm{t}_0)$, and taking the derivative, we obtain

$$\dot{M} = \dot{n}t + n - \dot{n}t_0 - n\dot{t}_0 = \dot{\varepsilon} - \dot{e}\sin\varepsilon - e\dot{\varepsilon}\cos\varepsilon \tag{3.89}$$

Thus,

$$\dot{t}_0 = (1/n)\,[-\dot{n}\,t_0 + n + \dot{n}t - \dot{\varepsilon}\,(1 - e\cos\varepsilon) + \dot{e}\sin\varepsilon]$$

or

$$\dot{t}_0 = (1/n)\,[\dot{n}\,(t - t_0) + n - \dot{\varepsilon}\,(1 - e\cos\varepsilon) + \dot{e}\sin\varepsilon] \tag{3.90}$$

Differentiating (3.40),

$$\begin{aligned}\dot{n} &= D\,(\mu/a^3)^{1/2} = 1/2(\mu/a^3)^{-1/2}(-3\mu a^{-4})\dot{a} \\ &= -(3/2)\left(\mu/a^3\right)\left(\dot{a}/a\right)/(\mu/a^3)^{1/2} = -\left(3/2\right)n\dot{a}/a\end{aligned} \tag{3.91}$$

and $\dot{a}$ is known from (3.68), so

$$\begin{aligned}\dot{n} = -(3/2)\left(n/a\right)&\left\{2a^{3/2}/\sqrt{\mu\,(1-e^2)}\right\} \\ \times &\left[(e\sin\upsilon)R + \left(1 + e\cos\upsilon\right)T\right]\end{aligned} \tag{3.92}$$

The quantity $\dot{\varepsilon}$ in (3.89) and (3.90) is found from (3.47), where the eccentric anomaly was represented by the symbol E. Rewriting it,

$$\dot{\varepsilon} = n/[1 - e\cos\varepsilon]$$

Thus the second and third terms of (3.90) cancel and, finally, $\dot{e}$ is known from (3.76). Substitution into (3.90), and rearranging, results in the following:

→

$$\begin{aligned}\dot{t}_0 = &\left\{\left\{-3\,(t-t_0)/\left[\mu\,(1-e^2)/a\right]^{1/2}\right\} e\sin\upsilon\right.\\ &+\left.\left[a^2(1-e^2)^{1/2}/\mu\right]\ \sin\upsilon\right\}R\\ &+\left\{\left\{-3\,(t-t_0)/\left[\mu\,(1-e^2)/a\right]^{1/2}\right\}[1+e\cos\upsilon]\right.\\ &+\left.\left[a^2(1-e^2)^{1/2}/\mu\right]\ [\cos\ \ \upsilon+\ \cos\varepsilon]\right\}T\end{aligned}\tag{3.93}$$

Note that only the components of the perturbing force in the plane of the orbit have a direct effect on the epoch. Note also that the effect grows with time because of the $(t - t_0)$ term on the right-hand side of (3.93). A more sophisticated treatment of the effect on the epoch can be found in Brouwer and Clemence (1961, pp. 285–289), Danby (1962, p. 243), or Murray and Dermott (1999/2001, p. 57).

For our purposes, it suffices that we have demonstrated how perturbations can affect the elements of the orbit. It is beyond the scope of the present work to discuss how to find the perturbing force given the change to the elements in general, but we will discuss later the determination of the distribution of mass in a planet given the perturbations of the orbit of a satellite. Other topics in, and techniques of, celestial mechanics can be obtained from the three excellent sources mentioned in the previous paragraph. As a practical interest, many programs (written in BASIC) are provided in Danby (1988).

This concludes our brief summary of celestial mechanics. We now turn to the physical components of the solar system, starting with its dominant member, the Sun.

Challenges

[3.1] Derive Kepler's third law from Newton's gravitational and motion laws. [Hint: Consider areal speed and make use of (2.37).]

[3.2] Suppose a projectile is shot vertically from a site on the equator with initial velocity less than escape velocity by, say, 1 km/s. Discuss what happens to the projectile.

[3.3] Assume that the initial orbit for an Earth satellite is 100 km above the Earth's (mean) equator and that its final orbital radius is that of a geo-synchronous satellite. (a) Compute the orbital elements and other parameters for the Earth satellite transfer orbit. Compute the velocities of (b) the circular orbits and (c) the transfer orbit at points of thrust and thus the velocity difference required to achieve the changed orbit. For these purposes, you can ignore other perturbations.

[3.4] (a) Compare the orbital properties of synchronous satellites on Mars and Earth and (b) compute the orbit of a synchronous satellite of the Moon. (c) Demonstrate the feasibility or non-feasibility of such a lunar satellite.

[3.5] There have been discussions about spacecraft paths that make full use of the gravitational attraction of solar system objects and thus minimize thrusts and use of chemical fuel. (a) Discuss the celestial mechanics involved in the design of such a 'highway' to the outer planets; (b) write down an expression for the net acceleration on the spacecraft at some instant; and (c) describe an iterative process which can be used to predict its future path without additional use of its rockets.

[3.6] If we now know the masses of all the planets to high precision, why is it difficult to predict the exact positions of Earth-crossing asteroids a few decades into the future?

[3.7] Derive an approximate expression for the maximum distance the Moon can recede from the Earth and retain its satellite status, and compute that distance. [Hint: Use (3.55) and set Δr equal to the distance between the Earth and the Moon. If all else fails, you can consult the more exact solution in Sect. 13.2 of Milone and Wilson (2014).]

References

Applegate, J.H., Douglas, M.R., Gursel, Y., Sussman, G.J., Wisdom, J.: The outer solar system for 200 million years. *Astron. J.* **92**, 176–194 (1986)

Bradstreet, D.H., Steelman, D.P.: Binary Maker 3.0*, Contact Software*, Norristown, PA (2004)

Brouwer, D., Clemence, G.M.: *Methods of Celestial Mechanics*. Academic Press, New York (1961)

Consolmagno, G.J., Schaefer, M.W.: *Worlds Apart: A Textbook in the Planetary Sciences*. Prentice Hall, Englewood Cliffs, NJ (1994)

Danby, J.M.A.: *Fundamentals of Celestial Mechanics*, 1st edn. McMillan, New York (1962) (Reprinted in 1964)

Danby, J.M.A.: *Fundamentals of Celestial Mechanics*, 2nd edn. Willmann-Bell, Richmond, VA (1988)

Laskar, J., Gastineau, M.: Existence of collisional trajectories of Mercury, Mars, and Venus with the Earth. *Nature* **459**, 817–819 (2009)

Milone, E.F., Wilson, W.J.F.: *Solar System Astrophysics: Planetary Atmospheres and the Outer Solar System*, 2nd edn. Springer, New York (2014)

Montenbruck, O.: *Practical Ephemeris Calculations*. Springer, New york (1989) (Tr. by Armstrong A.H. of Montenbruck's *Grundlagen der Ephemeridenrechnung* 3. Verlag Sterne und Weltraum Dr Vehrenberg GmbH, Munich (1987))

Moulton, F.R.: *An Introduction to Celestial Mechanics*, Second revised edn. MacMillan, New York (1914) 10th Printing (1958)

Murray, C.D., Dermott, S.F.: *Solar System Dynamics*. University Press, Cambridge (1999/2001).

Neutsch, W., Scherer, K.: *Celestial Mechanics: An Introduction to Classical and Contemporary Methods*. B.I. Wissenschaftsverlag, Mannheim (1992)

Newhall, X.X., Standish, E.M., Williams, J.G.: DE 102: a numerically integrated ephemeris of the moon and planets spanning forty-four centuries. *Astron. Astrophys.* **125**, 150–167 (1983)

Nobili, A.M.: Long term dynamics of the outer solar system: review of LONGSTOP project. In: The Few Body Problem, *Procs. of IAU Colloquium 96*. pp. 147–163 (1988)

Peirce, B.O.: *A Short Table of Integrals*, 4th edn. Ginn and Co., Boston, MA (1957)

Schlosser, W., Schmidt-Kaler, Th., Milone, E.F.: *Challenges of Astronomy: Hands-On Experiments for the Sky and Laboratory*. Springer-Verlag, New York (1991/1994)

Smart, W.H.: *Celestial Mechanics*. Wiley, New York (1953)

Sussman, G.J., Wisdom, J.: Numerical evidence that the motion of Pluto is chaotic. *Science* **241**, 433–437 (1988)

Szebehely, V.G.: *Adventures in Celestial Mechanics*. University of Texas Press, Austin, TX (1989)

Wisdom, J., Holman, M.: Symplectic maps for the N-body problem. *Astron. J.* **102**, 1528–1538 (1991)

Chapter 4
The Core of the Solar System: The Sun

The planets and other features of the solar system are, as this name implies, dominated by the Sun. Therefore, we begin by placing this primary component in context, describe its properties as a star, and discuss the relevant astrophysics required to gain some insight into its nature and importance.

4.1 The Solar Context

The role of the Sun in our solar system can be demonstrated in these empirical data:

- Visually the Sun is the solar system's dominant object; it is a star, a self-luminous body, powered by nuclear reactions in its core.
- The Sun is the most massive object in the solar system.
- Nevertheless, the solar system's angular momentum is concentrated in the planets, mainly Jupiter.
- The planets all revolve in the same direction (CCW as viewed from above the north ecliptic pole), the same direction as the rotation of the Sun.
- The orbits are roughly coplanar, except for those of Mercury and groups of small-body and "dwarf planet" objects, such as Ceres in the asteroid belt, and Pluto in the outer solar system.
- The orbits are roughly circular, except for those of Mercury and groups of small-body and "dwarf planet" objects, such as Pluto, in the outer solar system.
- There is a debris field (asteroid belt) and evidence for clouds of (cometary) debris beyond the orbit of Neptune.
- There is evidence for differentiation in chemical composition across the solar system.
- The Titius-Bode "law" suggests an underlying—although perhaps incompletely realized—principle or scheme for the spacing of the planets' (and the largest asteroid's) orbits, as noted in Chap. 1. See Nieto (1972) for an extensive discussion.

E.F. Milone and W.J.F. Wilson, *Solar System Astrophysics: Background Science and the Inner Solar System*, Astronomy and Astrophysics Library,
DOI 10.1007/978-1-4614-8848-4_4, © Springer Science+Business Media New York 2014

These points are discussed in other contexts in other chapters; for now it suffices to state that all theories of the origin of the solar system must take these observations into account. In our investigation, the properties of each group of solar system objects will be examined, and in the end we will try to see how our perception of the solar system fits into a more general context that is being formulated following the discoveries of planets around other stars. We begin at our local center, with the Sun.

The Sun was revered as a god for its life-sustaining warmth and its light long before its gravitational dominance was recognized. The apparent motion of the Sun along the ecliptic and the resulting seasonal variation in insolation and thus warmth was recognized at least 6 millennia ago, and probably much earlier, as noted in Chap. 1. The phenomenological effects of the Earth's revolution and rotation can be probed with the tools described in Chaps. 2 and 3. We discuss in this chapter the radiative properties of sunlight and how light interacts with matter. The explanation of the observable properties of the Sun requires a brief summary of radiation laws and the review of a number of definitions. The black body radiation laws will be described amid the descriptive properties of the Sun and other stars. Neither the Sun nor any other star is a perfect black body radiator, but it is often convenient to compare their properties to those of black bodies and to try to understand the causes of the differences. The light and particle emissions of the Sun are critical to the understanding of planetary properties and phenomena, and so these emissions and their variations must be described also.

4.2 The Sun as a Star in the Milky Way Galaxy

The average distance of the Sun from the Earth is 149,597,870,700(3) m (the number in parentheses is the uncertainty in units of the last decimal place) (Piteva and Standish 2009), and this value was adopted by the IAU (International Astronomical Union) in 2012 as the value of the *astronomical unit* (au) exactly. The distance is sometimes expressed as the solar parallax, the *mean equatorial horizontal parallax*, equal to the displacement shift the Sun would appear to undergo at mean distance, as viewed alternately from the center and from the limb (horizon) of Earth, using the equatorial radius as the baseline. The above value of the au and an equatorial radius for the Earth of 6,378,136.6(1) m yield a ratio of $\tan \alpha = 4.2635212 \times 10^{-5}$ rad or $\alpha = 8.794143$ arc-sec for the mean equatorial parallax. These and most of the other data in this section are taken from the 2012 Astronomical Almanac, or Cox (2000, p. 340ff) or Allen (1973, p. 161ff), with occasional updates from other sources.

The mass of the Sun can be determined from a formulation of Kepler's third law,[1] the length of the year, and the mean distance of the Earth from the Sun,

[1] Generally expressed as $G(\mathfrak{M}_\odot + \mathfrak{M})\ P^2 = 4\pi^2 a^3$, where $\mathfrak{M}$ is the mass of a planet. For the Earth's orbit, we may write $G(\mathfrak{M}_\odot + \mathfrak{M}_\oplus)\ P^2 = 4\pi^2 a^3$, where the gravitational constant, $G = 6.67384(80) \times 10^{-11}\ \mathrm{m}^3\ \mathrm{kg}^{-1}\ \mathrm{s}^{-2}$ (Sect. 3.1), $\mathfrak{M}_\odot$ and $\mathfrak{M}_\oplus$ are the masses of the Sun and Earth, the orbital period of the Earth, $P = 3.155815 \times 10^7$ s, the mass of the Earth, $\mathfrak{M}_\oplus = 5.9722(6) \times 10^{24}$ kg, and the semi-major axis of the Earth's orbit, $a = 1$ au, the value of which is given above.

$$\mathfrak{M}_{\odot} = 1.9884(2) \times 10^{30}\text{kg}$$

This mass represents 99.9 % of the total mass of the solar system.

The radius of the Sun can be determined by direct measurement of its angular size. The angular semi-diameter of the Sun in a narrow passband of red light ($\lambda = 0.800$ μm) was determined from a 6-year series of daily ground-based measurements to be $\alpha = 959\overset{''}{.}680(9) = 0.00465266(4)$ radian (Brown and Christensen-Dalsgaard 1998). The apparent solar radius is found, after adjusting the distance for the position of the observatory relative to the center of the Earth, and for the difference between the barycenter of the Earth-Sun system and the center of the Sun, to be

$$\begin{aligned}\mathfrak{R}_{\odot} &= r\alpha = 1.495936 \times 10^{11}\text{m} \times 4.65266(4) \times 10^{-3}\\ &= 6.96008(7) \times 10^{8}\text{m}\end{aligned}$$

This is slightly larger than the previously accepted value (6.9599×10^{8} m) for the solar radius. However, for many purposes the important radius is not the apparent edge of the Sun at a particular wavelength, but the radius at which the local temperature of the Sun is equal to the effective temperature (defined in Sect. 4.4.1). The correction requires a solar model. From the mean of the radii corrected with two such models, Brown and Christensen-Dalsgaard (1998) obtained for this definition of the solar radius,

$$\mathfrak{R}_{\odot} = 6.95508(26) \times 10^{8}\text{m}$$

The departure of the Sun's geometric figure from a sphere is small, but measureable. The *oblateness*, ϵ, is the difference between equatorial and polar radius in units of the equatorial radius. From highly precise measurements obtained in 1992 and 1994 with a balloon-borne instrument (Lydon and Sofia 1996), a weighted mean value of $\epsilon = 8.92(78) \times 10^{-6}$ is obtained. In a recent determination, Kuhn et al. (2012) used data from the Helioseismic and Magnetic Imager (HMI) aboard NASA's Solar Dynamics Observer (SDO) over an interval of 2 years. As interpreted by Gough (2012) the result is $\Delta_v = -7.56(40) \times 10^{-6}$, where $\Delta_v = \epsilon \;\; R_{eq}/<R>$ and $<R> = (R_{eq} + R_p)/2$. The improved precision allows Kuhn et al. to suggest that the result does not vary with the sunspot cycle, discussed below. The departure of the Sun from a sphere is smaller than expected from current models.

From the mass and the radius of the Sun, we find its mean density,

$$<\rho_{\odot}> = \mathfrak{M}_{\odot}/\left[(4/3)\pi\mathfrak{R}_{\odot}^{3}\right] \tag{4.1}$$

Adopting 6.960×10^{8} m for the solar radius and 1.9884×10^{30} kg for the mass, we obtain

$$<\rho_{\odot}> = 1,408\,\text{kg m}^{-3} = 1.408\,\text{g cm}^{-3}$$

and the gravitational acceleration at the solar radius $\mathfrak{R}_{\odot}$,

$$g_{\odot} = G\mathfrak{M}_{\odot}/\mathfrak{R}^2_{\odot} \tag{4.2}$$

so that $g_{\odot} = 273.96\ \text{m/s}^2 = 2.7396 \times 10^4\text{cm/s}^2$.

Another important physical property is the solar *angular momentum*,

$$\begin{aligned}\mathbf{L}_{\odot} &= \sum \mathfrak{M} \cdot \mathbf{v} \times \mathbf{r} = I \cdot \boldsymbol{\omega} \\ &= 1.96 \times 10^{48}\,\text{g cm}^2/\text{s} \\ &= 1.96 \times 10^{41}\,\text{kg m}^2/\text{s}\end{aligned} \tag{4.3}$$

in magnitude. Here $\mathfrak{M}$, **v**, **r**, and $\boldsymbol{\omega}$ are the mass, velocity due to rotation, the distance from the rotation axis, and angular velocity due to rotation, respectively. We have used a left-hand rule for the cross product (a right-hand rule would require $\mathbf{r} \times \mathbf{v}$), and the summation is taken over all solar atoms and ions! I is the *moment of inertia*,

$$I = K \cdot \mathfrak{M}\mathfrak{R}^2 \tag{4.4}$$

(Sect. 5.4.3), where K is a constant whose value depends on the distribution of density within the body. For a uniform sphere, $K = 2/5 = 0.4$.

$$\omega = 2\pi/P_{\text{rotn}} \tag{4.5}$$

is the mean *angular velocity* and P_{rotn} is the rotation period. We may approximate the value of K by using $\omega = 2.85 \times 10^{46}\ \text{rad s}^{-1}$, the angular velocity of the Sun at the equator; then from (4.3), $I = 6.9 \times 10^{46}\ \text{kg m}^2$, and from (4.4), $K = 0.071$. Such a small value of K indicates that the Sun is highly centrally compressed.

The value for $\mathbf{L}_{\odot}$ is based on the reasonable assumption that the internal rotation is the same as at the surface, as suggested by helioseismology data (Beatty and Chaikin 1990, p. 24). The overall angular momentum of the entire solar system is, however,

$$\text{L}_{\text{total}} = 3.148 \times 10^{50}\,\text{g cm}^2/\text{s}, \text{ so that } \text{L}_{\odot}/\text{L}_{\text{total}} \approx 0.005$$

Thus the angular momentum of the Sun makes up a negligible part of the total angular momentum. This is a vital clue of the solar system's origin.

The stars in the vicinity of the Sun, kinematically defining the *local standard of rest* (LSR), are moving through space in a circular orbit at a speed of ~220 km/s and are located about 8.5 Kpc from the center of the galaxy. These values were formally adopted by the IAU in 1985. The "solar motion" is the velocity of the Sun with respect to the LSR. The determined components of this motion in the directions away from the galactic center (u), in the direction of galactic rotation (v), and normal to the galactic plane positive toward the North Galactic Pole (NGP) (w) are:

$$\begin{aligned} u_\odot &= -9 \ \text{km/s} \\ v_\odot &= +12 \ \text{km/s} \\ w_\odot &= +7 \ \text{km/s} \end{aligned}$$

The net motion is 16.5 km/s, directed toward the position: $\ell = 53°$, $b = 25°$ in Hercules, a direction known as the solar apex. Note that the components indicate that the Sun is approaching the galactic center, is slightly faster at present than the appropriate orbital speed for a circular orbit at the current distance of the Sun, and also has motion normal to the plane. Thus the Sun has a non-circular and non-planar orbit about the galactic center. It is at the edge of a spiral arm and its basic properties identify it as a member of the "old disk" population. Its age is estimated at about 5×10^9 years.

4.3 Observable Properties of the Quiet Sun

The title of this section reveals that the Sun is a kind of Dr Jekyll and Mr Hyde. In this section we discuss its gentler nature, but the Mr Hyde aspect (the "active" Sun) is never more than half a solar (or "sunspot") cycle away. We briefly discuss the active Sun in Sect. 4.8. Direct imaging observations across the electromagnetic spectrum, in addition to eclipse observations, reveal that the Sun's outer regions consist of three parts:

The *photosphere* (literally "sphere of light")
The *chromosphere* ("sphere of color")
The *corona* or halo

The chromosphere and corona are best seen during total solar eclipses, either natural or artificially produced eclipses, with instruments known as *coronographs*, but the eclipse condition is not strictly necessary. Direct images of the Sun, taken with a type of spectrograph known as a *spectroheliograph*, in very narrow passbands, can be recorded as *spectroheliograms*, pictures of the Sun at higher regions of the atmosphere than can be seen in "white light" (very broad passband) images.

The motions of neutral and ionized atoms in the Sun can be studied through the analysis of the profiles of spectral lines. Contributions to the width of a spectral line come from several sources: (a) intrinsically, from the uncertainty principle; (b) from the relative abundance of the atoms in the particular ionization stage and excitation states involved in the atomic transition; and (c) the motion and the pressure in the atmosphere. Some of this motion arises from the kinetic energy of the atoms and is therefore dependent on the temperature. Other motions are due to the turbulence on small spatial scales (*microturbulence*) and on larger scales (*macroturbulence*) or the bulk motion of large areas of the Sun (*granulation*, *supergranulation*, and rotation). Spectrographs at higher spectral and spatial resolution are used to sort out these different effects. Microturbulence is invoked often as a last resort to account for the observed, but otherwise unexplained, widths of spectral lines.

The magnetic field structure of the Sun is studied both through the coronal structure and *magnetograms* produced by polarization-measuring spectrographs. Weak fields pervade the entire solar surface, but stronger concentrations are found in certain regions, especially those associated with so-called active regions, around sunspots.

Strongly correlated fluctuations on time scales of minutes and hours in the brightness and velocity of many regions over the solar disk have begun to reveal details about the interior structure of the Sun through the techniques of helioseismology.

Finally, particle emissions of the Sun are observed with detectors on satellites and probes and even indirectly on the Earth with cosmic ray detectors and through other indirect effects on the ionosphere and upper atmosphere. One particle, the neutrino, produced in nuclear reactions in the core, has been of great significance in the past few decades. The neutrino observations revealed a shortfall of about a factor of 3 in neutrino flux produced from expected nuclear reactions in the proton–proton chain that is the expected source of solar energy. Near the end of the twentieth century, this neutrino problem was one of the major unsolved problems of astrophysics. The solution lay in the oscillation between different forms of the neutrino, from that related to the electron to those related to the tau meson and the mu meson (the tauon and muon, respectively). Earlier experiments were sensitive only to the electron neutrino, and so failed to see the number that converted to the other “flavors” by the time they reached the detectors. The critical evidence for neutrino oscillation is presented in Ahmad et al. (2001) and an examination of the agreement among solar models, solar seismology observations, and neutrino data can be found in Bahcall et al. (2005).

4.4 The Sun’s Radiation

In order to discuss the radiative properties of the Sun, a few general ideas, definitions, and distinctions need to be considered.

4.4.1 Luminosity and Surface Brightness

In spherically symmetric stars, the radiant power emitted to the rest of the universe (the *luminosity*) is given by

$$\mathcal{L} = 4\pi \cdot \mathfrak{R}^2 \mathfrak{S}(\text{watts or W}) \tag{4.6}$$

where $\mathfrak{R}$ is the stellar radius and $\mathfrak{S}$ (a script S) is the power radiated per unit surface area. In some branches of astronomy, $\mathfrak{S}$ [sometimes written as an unscripted S, a script B, σ, $\sum$, or, because it has units of flux, as $\mathscr{F}$] is called the *surface flux*, or, sometimes, the *surface brightness*. It may also be called the *radiance*, which Dufay (1964, p. 5) defines as "the flux radiated by a unit surface element in all exterior directions."[2] Most often, the quantity referred to as the surface brightness or radiance is expressed in units of power per unit area per unit solid angle of emission (see further discussion below). With these units, these quantities measure intensity. The distinction is important and will be discussed further in this section and in Sects. 4.4.2 and 4.4.3.

In (4.6), radiation has been integrated over the entire sphere. In any case, $\mathfrak{S}$ depends on the temperature, which is basically a measure of the average kinetic energy of a large number of particles. $\mathfrak{S}$ may be written

$$\mathfrak{S} = \sigma T^4 \quad \left(\text{W/m}^3\right) \tag{4.7}$$

where σ is the *Stefan-Boltzmann constant*, $(5.6705 \pm 2) \times 10^{-8}$ W m^{-2} K^{-4}, and T is called the *effective temperature*—the temperature at the star's surface in order to have a surface brightness equal to that of the idealized, perfect radiator that we commonly call a *black body* (BB). Indeed, the surface of a star is *not* a black body, as comparisons between BB curves for particular temperatures and real star's radiation curve tracings readily reveal (see Fig. 4.1). However, for many purposes, especially in the interiors of stars, BB approximations are very useful.

Before we discuss black body radiation, however, we must review the physical terms we have just used and discuss how these terms fit into astronomy, and the more specific fields of practical physics called radiometry and photometry.

The *luminosity*, already defined in (4.6) and the text near it, as the amount of energy radiated per second at the source (in the cases described here, the Sun or a star) is also called the *radiant power* or, sometimes, regrettably, "*radiant flux*" (in units of W). The power radiated only in a spectral region centered around the wavelength 0.555 μm (555 nm, 5,550 Å), which is the approximate wavelength of peak sensitivity of the human eye in daylight, is referred to as *luminous power*, or, sometimes as "*luminous flux*." The unit of luminous power is, naturally enough, the *lumen* (abbreviated lm), equivalent to[3] 1/683 W.

[2] If one is discussing the measurement of light from a source, *irradiance* is a term sometimes used to describe the radiation in a wide passband that falls onto a surface (of a telescope, or a detector, for instance); in the visual region, the term is *illuminance*. If the measurement involves such received radiation per unit of wavelength, it is called *spectral irradiance*.

[3] See Meyer-Arendt (1995, p. 351); the sensitivity of the human eye to different wavelengths makes this conversion factor vary somewhat with wavelength and bandwidth.

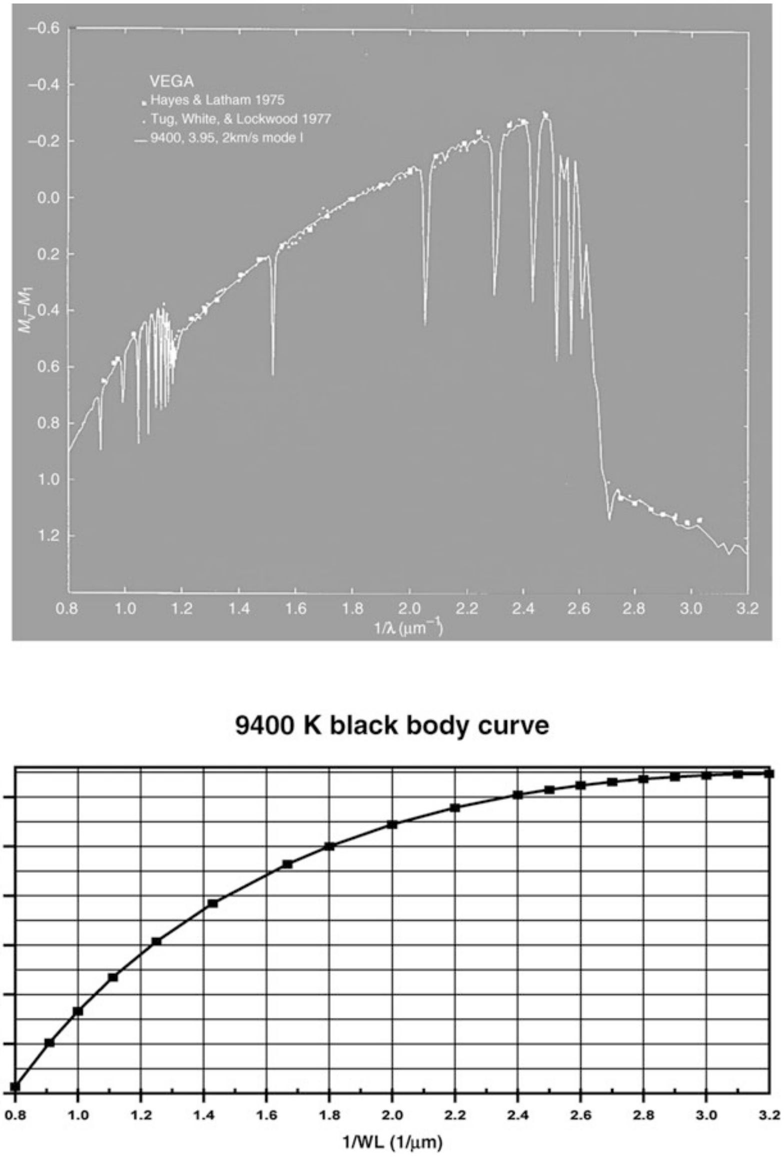


Fig. 4.1 (**a**) From Kurucz (1993), the spectral irradiance of Vega, relative to its value at 555.6 nm and converted to magnitudes, compared to Kurucz's model and others. (**b**) The monochromatic flux of a 9,400 K black body, the effective temperature of that model, and normalized and converted to magnitudes as in (**a**). The abscissa is in units of inverse wavelength, so that a wavelength of 0.5 μm is located at 2.0 (μm)$^{-1}$

Radiant exitance (*luminous exitance* for the visual region) is the power emitted per unit area in units of W/m^2 (and in the visual, lm/m^2); *radiant intensity* (*luminous intensity* for the visual region) is the amount of *radiant* power (*luminous* power in the visual) emitted per unit solid angle,[4] Ω, and has units of W/sr (for the visual region, this is in units of lm/sr). If a source emitting monochromatic radiation at a frequency of 540×10^{12} Hz (i.e., at a wavelength of 555 nm) into a given direction has a radiant intensity of (1/683) W in that direction, the luminous intensity would be 1 lm/sr, a quantity also called a *candela* (cd).

Finally, the radiant power passing through a unit area and into a unit solid angle at the source is the *radiance*; *luminance* refers to the visual component of the radiance. Radiance and luminance are used to describe the power emitted at different regions of the emitter's surface, and they are sometimes referred to as "brightness," or "surface brightness." The units are W m^{-2} sr^{-1} for radiance, and cd/m^2 [or *nit* (for the Latin *nitere*, "to shine")], for luminance. Alternatively the luminance is given in units of lamberts ($10^4/\pi$ cd/m^2). In astronomy, the relative brightness is often expressed in magnitudes, related to the various units of brightness through the base 10 logarithm of the brightness ratio: $(m - m_0) = -2.5 \log (\ell/\ell_0)$. The only requirement is that ℓ and ℓ_0 must be measured in the same units. We will discuss magnitudes and their uses in a later section and in later chapters.

In Fig. 4.1, we show the monochromatic (i.e., per unit wavelength) flux in the form of spectral irradiance of the star Vega, along with predicted values from several models (cited in Kurucz 1993) and, separately, in the same units, the corresponding radiation curve of a black body at 9,400 K, the temperature assumed for the model. The spectral irradiance is normalized to (i.e., divided by) its value at 555.6 nm = 0.5556 μm equivalent to the inverse wavelength, 1.8 (μm)$^{-1}$, against which units the spectra are plotted. Vega's spectrum shows absorption features at individual wavelengths (*called spectral lines*) and in the continuum. The image is reversed to white on black to emphasize this point.

The role of opacity in shaping spectral lines will be featured in Sect. 4.9. Absorptions over a large wavelength range are due to *bound–free* and *free–free transitions*. The former can be caused by ionizations of atoms by photons with energies greater than the ionization energy of the atom. The edge of the Balmer series of hydrogen, caused by ionization of the hydrogen atom with electrons in the second energy level, is seen at ~0.365 μm, equivalent to 2.74 (μm)$^{-1}$. Radiation streaming through the star's atmosphere that is shorter in wavelength (bluer) than

[4] Angles are measured in degrees or radians (2π radians = 360°). Solid angles are measured in square degrees or steradians (sr). Generally, the solid angle in sr taken up by an area on a sphere when viewed from the center is Ω = area/(radius)^2. 1 steradian is thus the angle subtended by an area of one square meter on the surface of a sphere of 1 m radius (NB: the area can be of any shape). The surface area of a sphere of radius R meters is $4\pi R^2$ m^2, so the solid angle taken up by the complete sphere as viewed from the center is $\Omega = 4\pi$ sr; Also, $(\pi/180)^2 = 3.0462 \times 10^{-3}$ sr/deg^2 and 1 sr = $(180/\pi)^2 = (57.296)^2 = 3282.8$ deg^2; the entire sphere subtends at the center 4π sr = 41, 252.88 deg^2.

the *Balmer limit* is readily absorbed by atoms in this energy state, so that the outward going flux is greatly reduced.

4.4.2 Flux and Intensity

Four important quantities used to describe an object's radiative properties are:

1. The emittance, surface brightness (as we have referred to it here), or more commonly, *flux*, F (or $\mathscr{F}$), from an object is the total radiant energy emitted per second per unit area (W/m^2) in all outward directions (2π sr).
2. The *specific flux*, F_ν or F_λ, is the flux per unit f/wl range. The frequency, f, is usually expressed in hertz (Hz) or cycles per second, but may be also in kilohertz (kHz), megahertz (MHz), gigahertz (GHz), etc., whereas the wavelength, "wl" in our shorthand, may be in units of m, cm, mm, nm, μm, Ångström units (Å) = 10^{-10} m, etc. Thus the specific flux may be expressed in units of W/m^2/Hz or W/m^2/wl unit.
3. The *intensity* is a directed flux of energy, i.e., energy per second per steradian in a particular direction, per unit area normal to the radiant direction (W/m^2/sr).
4. The s*pecific intensity*, I_ν or I_λ, is the intensity per unit f/wl range (W/m^2/sr/Hz or W/m^2/sr/wl unit).

The important distinction between flux and intensity is that the intensity depends on the angle of emergence from a glowing object.

4.4.3 Black Body Radiation

In local thermodynamic equilibrium[5] (usually abbreviated to LTE), or for radiation emitted by any black body, $I_\nu = B_\nu(T)$ and $I_\lambda = B_\lambda(T)$, where B_ν and B_λ are the *Planck function* (sometimes called the *Kirchhoff-Planck function* or the *Planckian*):

[5] *Thermodynamic equilibrium* is an idealized state in which all processes are balanced in detail by their inverses; therefore, there can be no net transfer of radiation in any direction, and the temperature must be uniform everywhere. The interior of the Sun or any other star cannot be in thermodynamic equilibrium, because temperature decreases monotonically outward from the center to the surface, and consequently there is a net transfer of radiation in the outward radial direction at all points. However, except close to the photosphere, material in the stellar interior is sufficiently opaque that the mean free path for photons is of the order of centimeters or less, and over such a small distance the temperature and the radiation field have the appearance of being almost completely uniform. In this case it is permissible, to very high accuracy, to use the equations of thermodynamic equilibrium. Such a state is referred to as *local thermodynamic equilibrium.*

$$\begin{aligned} B_\nu(T) &= \left(2h\nu^3/c^2\right)\left(e^{h\nu/kT}-1\right)^{-1} \quad \left(\mathrm{Wm^{-2}Hz^{-1}}\right) \\ B_\lambda(T) &= \left(2hc^2/\lambda^5\right)\left(e^{h\nu/kT}-1\right)^{-1} \quad \left(\mathrm{W/m^2/wl\ \ unit}\right) \end{aligned} \tag{4.8}$$

where Planck's constant, $h = 6.6261 \times 10^{-34}$ J s, the speed of light, c = 2.9979 × 10^8 m/s, and Boltzmann's constant, $k = 1.3807 \times 10^{-23}$ J/K.

The setting of the derivative of (4.8) to zero defines the peak of the curve, and the f/wl at which this occurs is related to T through:

$$\begin{aligned} \nu_{\max} &= 0.5879T \ \ (Hz) \\ \lambda_{\max} &= 0.002898/T \ \ (m) \end{aligned} \tag{4.9}$$

with T in kelvins (K). The latter expression of (4.9) is known as *Wien's law*.

The quantity B_ν or B_λ has units of specific intensity and, in the present context, is sometimes called the *source function*. Allen (1973, p. 104) defines the integration of this quantity over all f/wl as:

$$B(T) = (\sigma/\pi)T^4 = 1.80468 \times 10^{-5}\,T^4\left(\mathrm{erg\,cm^{-2}\,s^{-1}\,sr^{-1}}\right) \tag{4.10}$$

evaluated in CGS units, i.e., B is the emittance per unit solid angle (steradian). The constant $\sigma = (\pi^2/60)k^4/[(h/2\pi)^3c^2] = 5.6705 \times 10^{-8}$ Wm^{-2} K^{-4} is often called the *Stefan-Boltzmann constant*, and T is not only the effective temperature, but also the local temperature as measured at any point on the black body (a consequence of LTE—see fn 5 and Sect. 4.9). Equation (4.10) is called the *Stefan-Boltzmann law*, and so is (4.7) because the temperature there is understood to be an *effective* temperature.

The flux can now be found in terms of the Planck function by integrating (4.10) over all outgoing (increasing r) directions:

$$\begin{aligned} \mathscr{F} &= \int B(T)\cos\theta\,\mathrm{d}\omega = \int_{\varphi=0}^{2\pi}\int_{\theta=0}^{\pi/2} B(T)\cos\theta\sin\theta\,\mathrm{d}\theta\,\mathrm{d}\varphi = 2\pi B(T)\int_{\theta=0}^{\pi/2}\sin\theta\cos\theta\,\mathrm{d}\theta \\ &= 2\pi B(T)\int_{\theta=0}^{\pi/2}\sin\theta\,\mathrm{d}(\sin\theta) = 2\pi B(T)\left[\tfrac{1}{2}\sin^2\theta\right]_{\theta=0}^{\pi/2} = \pi B(T) \end{aligned} \tag{4.11}$$

where θ is the angle with respect to the surface normal of a particular beam of radiation. Figure 4.2 illustrates the geometry, where we now apply these concepts to a star. At the center of a stellar disk the intensity is I_0, and at some point where the line of sight is at an angle θ, the intensity is often given as:

$$I = I_0\,(1 - u + u \ \ \cos \ \ \theta) \tag{4.12}$$

where u is the linear limb-darkening coefficient. When u = 0, the disk is uniformly bright, and when u = 1, the limb is fully darkened. See (4.32) and the discussion of *limb-darkening* in Sects. 4.5.1 and 4.5.2.

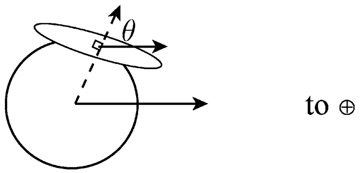


Fig. 4.2 Directed beams of radiative intensity

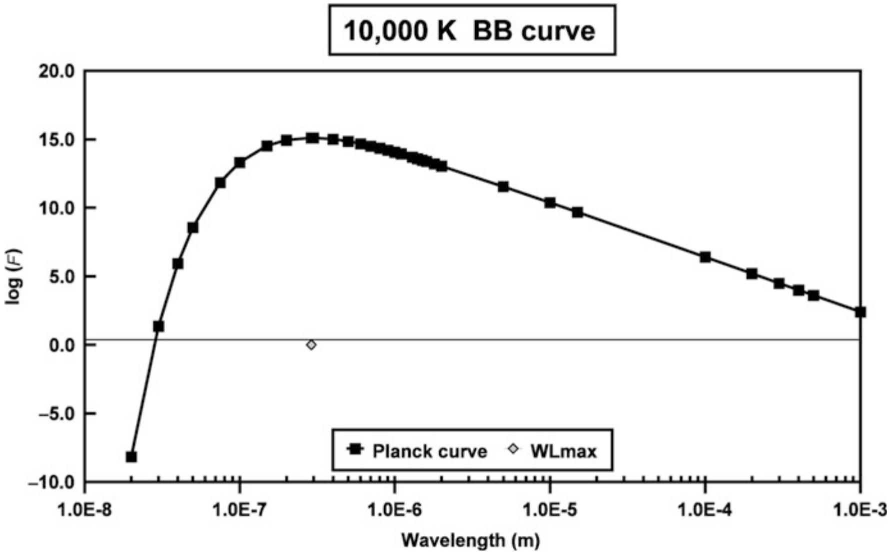


Fig. 4.3 The Planck function, πB_λ, for a black body at $T = 10{,}000$ K. Note that the peak of the radiation curve occurs near $\lambda = 3 \times 10^{-7}$ m, marked by the lozenge on the line parallel to the x-axis

From (4.11) and (4.12), the flux from a black body is

$$\mathscr{F} = \sigma T^4 \left(\mathrm{W/m^2}\right) \tag{4.13}$$

Note that the quantity that we have called $\mathfrak{S}$ in (4.6) and (4.7) (and written as a script B sometimes in other sources) and F (or $\mathscr{F}$) in (4.11) and (4.13) have units of power per unit area, whereas the B of (4.10) has units of power per unit area per steradian of solid angle. These are the units of *intensity*.

In Sect. 4.9 we discuss the transport of radiation through material with a "transfer equation" and use it to explain how emission and absorption features arise in the spectra.

Example 4.1

Figure 4.3 is a plot of the base 10 logarithm of the Planck function for a 10,000 K black body, computed from the second part of (4.8) against the base 10 logarithm of the wavelength. The peak of the curve, computed from the second

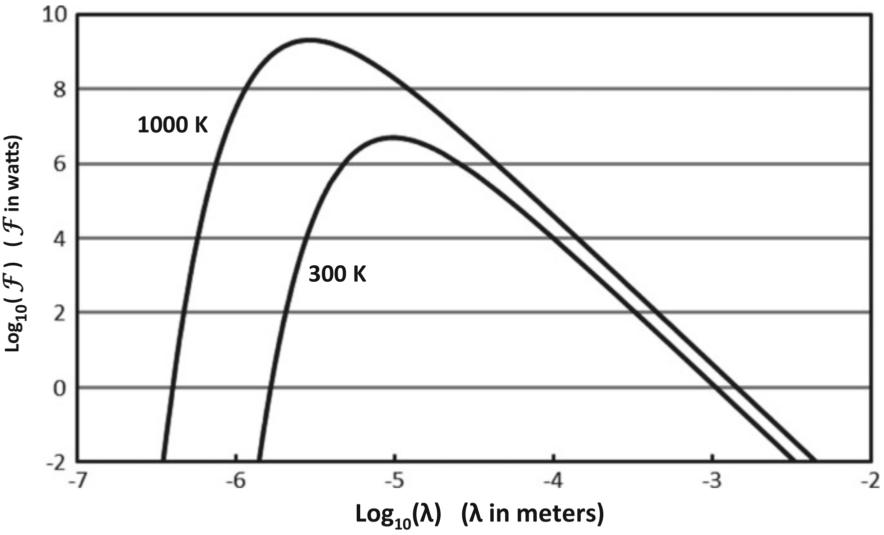


Fig. 4.4 The Planck functions for black bodies at $T = 1{,}000$ and 300 K. Note the shift to the blue of the wavelength of the peak of the higher temperature radiation curve, and that the lower temperature curve completely nests within the higher temperature curve

part of equation (4.9), is: $\lambda_{\max} = 2{,}898/T$ (μm) $= 0.2898$ μm $= 289.8$ nm, in the ultraviolet (UV).

Example 4.2
The black body curves computed for different temperatures are nested: lower temperature curves fall completely below higher temperature curves at each and every f/wl. Figure 4.4 illustrates this for Planck functions of two different temperatures, 1,000 and 300 K, respectively. This means that the flux integrated across all f/wl will also be greater for a higher temperature black body, as predicted by the Stefan-Boltzmann law (4.10). Note also that the higher temperature black body would appear bluer because the peak of its radiation curve occurs at shorter wavelengths.

4.4.4 Observed Radiative Properties of the Sun

The *solar constant* is the flux of total radiation received outside the Earth's atmosphere per unit area at the mean distance from the Sun. It is equal to

$$\begin{aligned}\mathfrak{C} &= 1.950(4)\ \mathrm{cal\,cm^{-2}\,min^{-1}} = 1.360\times10^{6}\ \mathrm{erg\,cm^{-2}\,s^{-1}}\\ &= 1.360\times10^{3}\ \mathrm{W\,m^{-2}}\end{aligned}$$

(1 cal = 1 calorie = 4.184 × 10^7 erg = 4.184 J), where we have used a script C to represent the solar constant. The luminosity can then be determined by multiplying the solar constant by the total surface area of a sphere of radius 1 au:

$$\begin{aligned}\mathcal{L}_\odot &= 4\pi r^2 \mathfrak{C} \\ &= 3.826(8) \times 10^{33}\ \text{erg/s} = 3.826(8) \times 10^{26}\ \text{W}\end{aligned} \tag{4.14}$$

This is one way to determine the luminosity. Another way is to multiply the surface area of the Sun by its surface brightness. From these data, the mean solar surface brightness and the mean solar intensity can be determined:

$$\mathscr{F} = \mathfrak{S} = 6.27 \times 10^7 \text{W/m}^2 = 6.27 \times 10^7 \text{erg cm}^{-2}\text{s}^{-1}$$

$$I = \mathscr{F}/\pi \tag{4.15}$$

So, $I = 2.000 \times 10^{10} \text{erg cm}^{-2}\text{s}^{-1}\text{sr}^{-1} = 2.000 \times 10^7 \text{W m}^{-2}\text{sr}^{-1}$

Once the surface brightness, $\mathfrak{S}$, is known, the effective temperature can be computed from (4.13). One such determination is T_{eff} = 5,777 K (W. C. Livingston, in Cox 2000, p. 341).

The observed visual magnitude of the Sun is $V = -26.75$ and its color index (B − V) = 0.65. The absolute magnitude, M, is the magnitude of a star as it would appear at a distance of 10 pc. Therefore,

$$M = m - 5\ \log(r/10) \tag{4.16}$$

where r is expressed in parsecs. From (4.16), $M_{\odot V} = 4.82$. The *bolometric correction* (BC) is the difference between the visual and bolometric magnitudes. (The latter is the magnitude of the flux of the star over all wavelengths.) Defined as in (4.17), if non-zero, the BC is almost always *negative*. However, because of differing zero points, BC values for some stars in some tables may be positive (Torres 2010).

$$M_{\text{bol}} = M_V + BC \tag{4.17}$$

As formulated here, the bolometric correction for the Sun is −0.08 so its bolometric absolute magnitude is 4.74. Slightly different values may be obtained elsewhere in the literature; we used the data given in Allen's Astrophysical Quantities (Cox 2000, p. 341).

4.5 The Photosphere

The visible surface of the Sun is known as the photosphere (see Fig. 4.5). We see it through the overlying chromosphere and corona. The total thickness of the photosphere is only a few hundred kilometers, yet most of the visible radiation comes

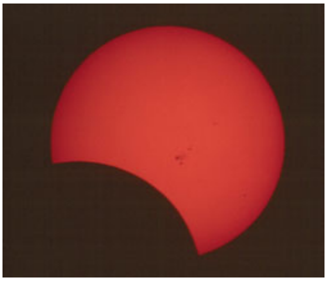


Fig. 4.5 The photospheric disk of the partially eclipsed Sun, February 26, 1979, showing limb-darkening and sunspots. Courtesy Dr. T. A. Clark

from this region. The reason for this interesting circumstance can be found in an understanding of opacity and limb-darkening, which we discuss in the following section. The photosphere marks the upper end of the convection zone, the convective effects of which are manifested in granulation patterns on the photosphere. We will discuss first the visibility of radiation and the causes of limb-darkening, then the convection zone, and finally other features of the photosphere.

4.5.1 Opacity and Optical Depth

We start by defining the unitless *optical depth*, τ, a measure of the extent to which radiation can penetrate into any medium. The contribution to the optical depth of a cylinder of gas of physical length *dx*, cross-section per unit mass or mass absorption coefficient[6] k_λ (expressed in units of m^2/kg), and density ρ (in units of kg/m^3) is:

$$d\tau_\lambda = k_\lambda \rho \, dx \tag{4.18}$$

Here, λ stands for wavelength, so k_λ and τ_λ are appropriate for light of that wavelength.

Integrating over a total physical distance $x = 0$ to X, we get:

$$\tau_\lambda = \int_{x=0}^{X} k_\lambda \rho \, dx \tag{4.19}$$

When $k_\lambda \rho$ is large and/or the physical length is great, τ_λ is also large and so is the absorption. Note that it is the product of the physical length and the absorption coefficient $k_\lambda \rho$ that determines the optical depth and thus the absorption. A very long path length need not guarantee strong absorption (or emission) if the opacity or the density is very small. If the density and opacity are constant over this total length, one could write

[6] Note that the *absorption coefficient* discussed in, e.g., Schlosser et al. (1991/4) is in units of inverse length, m^{-1}, not in m^2/kg, the units for the mass absorption coefficient, k_λ. This is because we show the density dependence explicitly here; but in (4.34), Sect. 4.9, we define and use the absorption coefficient. Generally, the absorption coefficient is referred to as *opacity*.

$$\tau_\lambda = k_\lambda \rho (X - 0) = k_\lambda \rho X \tag{4.20}$$

With the optical depth defined, we now consider the transport of radiation through the solar atmosphere. When radiation travels radially outward through a star, the intensity of the radiation changes over any infinitesimal distance, dr, due to the opacity of the material through which it travels

$$\mathrm{d}I_\lambda = -I_\lambda k_\lambda \rho \,\mathrm{d}r = -I_\lambda \mathrm{d}\tau_\lambda \tag{4.21}$$

where I_λ is the monochromatic (one-wavelength) intensity at position r. The negative sign indicates a loss of intensity.

If we consider a pencil beam of radiation within the star, and we know the intensity, $I_{\lambda,0}$, at some point in this beam, then we would like to find the intensity, I_λ, after the radiation has travelled from there to some later point in the beam. The relationship can be found by dividing both sides of (4.21) by I_λ and integrating over the optical depth:

$$I_\lambda = I_{\lambda,0} \exp(-\tau_\lambda) \tag{4.22}$$

For example, (4.22) gives the intensity, I_λ, emerging from a star if we know the intensity, $I_{\lambda,0}$, at some depth inside the star.

At this point, it is important to be aware of a potentially confusing (but nevertheless very useful) inconsistency in terminology. Equation (4.22) requires that $\tau_\lambda = 0$ at the point inside the star where $I_\lambda = I_{\lambda,0}$, and τ_λ *increases outward* to a value $\tau_\lambda > 0$ at the stellar surface. However, we are outside the star looking in rather than inside the star looking out! Therefore, we must shift our vantage point. Imagine that we are able to place a bright light source at the surface of the star. The light will not be dimmed in travelling from the source to us, so the optical depth between the source and us is 0 for this path. Now if we gradually move the source deeper and deeper into the star, we will see that it appears fainter and fainter as the optical depth between the source and us becomes larger and larger. It is therefore often convenient to talk of optical depth as being zero at the stellar surface and *increasing inward* into the star. A point at the surface is at optical depth zero, and a point at optical depth 1 is just barely visible.

Of course radiation is not only absorbed and scattered out of the beam, but can be scattered and reemitted *into* it also. We can express the radiation injected into the beam at any point in terms of the *emission coefficient* per unit length, ε_λ:

$$\mathrm{d}I_\lambda = \varepsilon_\lambda \mathrm{d}r. \tag{4.23}$$

If the point is at an optical depth τ_λ, then only a fraction $e^{-\tau_\lambda}$ of the upwelling monochromatic radiation contributes to the radiation leaving the star. In Sect. 4.9, we will show that ε_λ can be expressed in terms of the Planck function, $B_\lambda(T)$, and the mass absorption coefficient, k_λ, by

$$\varepsilon_\lambda = B_\lambda(T) k_\lambda \rho. \tag{4.24}$$

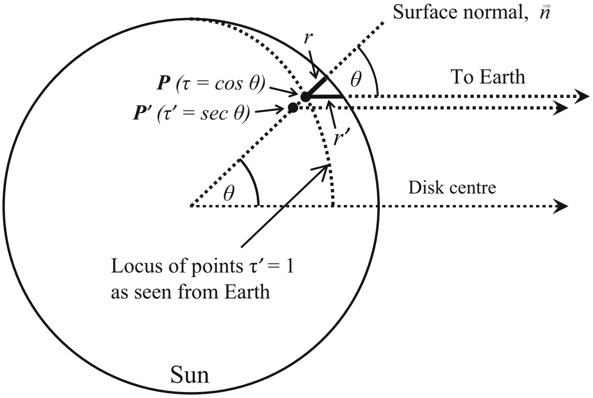


Fig. 4.6 Solar limb-darkening definitions and geometry. The physical depth of the $\tau' = 1$ locus is exaggerated for clarity

Then we can substitute (4.24) into (4.23) to find the contribution by the radiant energy emitted at any point in the star to the total radiation leaving the star is

$$\mathrm{d}I_\lambda = B_\lambda(T)\,k_\lambda \rho\, e^{-\tau_\lambda}\mathrm{d}r = B_\lambda(T)\,e^{-\tau_\lambda}\mathrm{d}\tau_\lambda \tag{4.25}$$

B_λ, r, τ_λ, and ε_λ increase *inward* (i.e., into the solar atmosphere from outside). T can be expressed as a function of τ_λ, so the total energy leaving the star is the sum of all contributions given by (4.25) from 0 to infinite optical depth:

$$I_\lambda = \int_0^\infty B_\lambda(\tau_\lambda)\,e^{-\tau_\lambda}\mathrm{d}\tau_\lambda \tag{4.26}$$

Upwelling radiation that starts out being well represented by a Planckian will undergo partial absorption due to the presence of absorbers in the line of sight.

From (4.22), along any particular ray, an optical depth of unity:

$$\tau_\lambda = 1 \tag{4.27}$$

corresponds to the depth over which the radiation falls to $1/e$, or 36.8%, of its original value.

From the outside, we see little radiation from depths deeper than $\tau = 1$. Therefore, this is a typical depth to which we can see into the atmosphere along any line of sight.

Note from Fig. 4.6 that from a given physical depth, say point P′, measured along a radius, the path length along a normal to a point on the surface is shorter than the path length through the solar atmosphere to the direction of the observer, unless the beam is emerging from the disk center. Thus the radiation is diminished over this "slant ray" distance compared to one traversing the "normal ray" distance. However, essentially we do not "see" deeper than $\tau = 1$ along any ray.

4.5.2 Center-to-Limb Variation

In Fig. 4.6, any point P is located at depth r and optical depth

$$\tau_\lambda = \int_{r=0}^{r=r} k_\lambda(r)\rho(r)dr \tag{4.28}$$

as seen by an observer viewing P along the surface normal.

An observer on the Earth, however, views P along a ray slanted at an angle, θ, from the normal, and thus at depth r′ and optical depth

$$\tau'_\lambda = \int_{r=0}^{r=r} k_\lambda(r')\rho(r')dr' \tag{4.29}$$

(For this section, we define r and r′ as depth variables, with r, r′ = 0 at the stellar surface.) We would like to find τ' in terms of τ.

From the geometry in Fig. 4.6, as we increment the depth along the slant ray by an amount dr′, the depth below the surface (along a surface normal) increases by only dr ≈ dr′ cos θ. (The approximation arises because the angle, θ, changes with depth along the slant ray; but if we limit ourselves to optical depths less than ~1, then r′ ≪ R, where R is the solar radius, and θ ~ constant.) Then

$$\mathrm{d}r' \approx \mathrm{d}r/\cos\theta = \sec\theta\,\mathrm{d}r \tag{4.30}$$

At each step in the integration in (4.29), k_λ is being evaluated at the same physical point located at depth r′ on the slant ray and depth r on the normal ray; thus, $k_\lambda(r') = k_\lambda(r)$. Similarly, $\rho(r') = \rho(r)$. Equation (4.29) then gives

$$\begin{aligned}\tau'_\lambda &= \int_{r'=0}^{r'=r'} k_\lambda(r')\rho(r')dr' = \int_{r=0}^{r=r} k_\lambda(r)\rho(r)\quad \sec\theta\, dr \\ &= \sec\theta \int_{r=0}^{r=r} k_\lambda(r)\rho(r)\quad dr = \tau_\lambda \sec\theta\end{aligned} \tag{4.31}$$

Looking into the Sun along the slant ray, we see to point P at optical depth $\tau'_\lambda = 1$; but from (4.31), this point would be at an optical depth of only

$$\tau_\lambda = 1/\sec\theta = \cos\theta \tag{4.32}$$

if viewed along the normal. Conversely, a point that would be at optical depth $\tau_\lambda = 1$ (point P′ in Fig. 4.6) when viewed along the normal to the surface is actually seen at optical depth $\tau'_\lambda = 1 \cdot \sec\,\theta = \sec\,\theta$ when viewed along the slant ray. Thus, by examining the intensity emerging from a star at a range of angles θ from the normal (i.e., from the direction of the center of the disk), a table of τ_λ vs. I_λ can be obtained.

If temperature decreases outward, as it does in the solar photosphere, the material at $\tau'_\lambda = 1$ is progressively cooler as we look at larger angles from disk centre, and I_λ decreases away from the centre of the disk. This is the *center-to-limb variation* or *limb-darkening* which can be seen in images of the Sun's disk. The limb-darkening is usually defined by the ratio

$$I_\lambda(\theta)/I_\lambda(0)$$

where $I_\lambda(\theta)$ is the intensity of the solar radiation arising from a point on the solar disk which makes an angle θ measured between a line from the center of the Sun to the point and a line from the Sun's center to the observer (as in Fig. 4.6).

The limb-darkening is often expressed analytically in the linear (but not always most accurate) form,

$$\begin{aligned} I_\lambda(\theta)/I_\lambda\big(0\big) &= 1 - u + u\cos\theta \\ &= 1 + u\left(\cos\theta - 1\right) \end{aligned} \tag{4.33}$$

where u is known as the *linear limb-darkening coefficient*. Therefore, for a fully darkened limb, $u = 1$; for no limb-darkening, $u = 0$. See Schlosser et al. (1991/4, Fig. 28.7) for a plot of $I(\theta)/I(0)$ for two passbands and for a plot of T vs. r (Fig. 28.8) in the Sun. Limb-darkening can be seen even in the darkened limb of white light images of the Sun. It varies slowly with wavelength in the continuum, but generally differs from one spectral line to another, because of the different opacities per absorbing element and the different conditions at the atmospheric heights where the atoms contributing to the line are located. Figure 4.7 shows the intensities of a star similar in temperature to, but much larger than, the Sun at different limb positions ($\mu = \cos\,\theta$). The significance of limb-darkening lies in the circumstance that the temperature increases with depth into the star.

Assuming a value of T_0, the temperature at the center of the disk (not at the center of the Sun!), one can then determine a series of values of T vs. τ and thus provide information about the thermal structure of the solar atmosphere. For the Sun, $I_0 = 2.41 \times 10^7(\mathrm{W\ m^{-2}\ sr^{-1}})$, which yields a disk-center temperature, $T_0 \approx 6{,}050$ K.

Note that if the temperature were to decrease with depth, we would expect (depending on the behavior of the opacity) greater emission from the limbs than from the disk center. This is known as *limb-brightening*, and this is actually seen in the Sun at some wavelengths where the emission arises in the Sun's upper atmospheric levels. The more general term, therefore, is *center-to-limb variation*.

In the solar photosphere, the temperature increases inward. The temperature minimum is reached near the boundary between the photosphere and the overlying chromosphere, through which it climbs to very high values.

In the ultraviolet, the center-to-limb variation shows more complicated behavior than the linear form of (4.33) predicts (we refer to the SkyLab satellite analyses of solar UV center-to-limb variation by Kjeldseth Moe and Milone 1978).

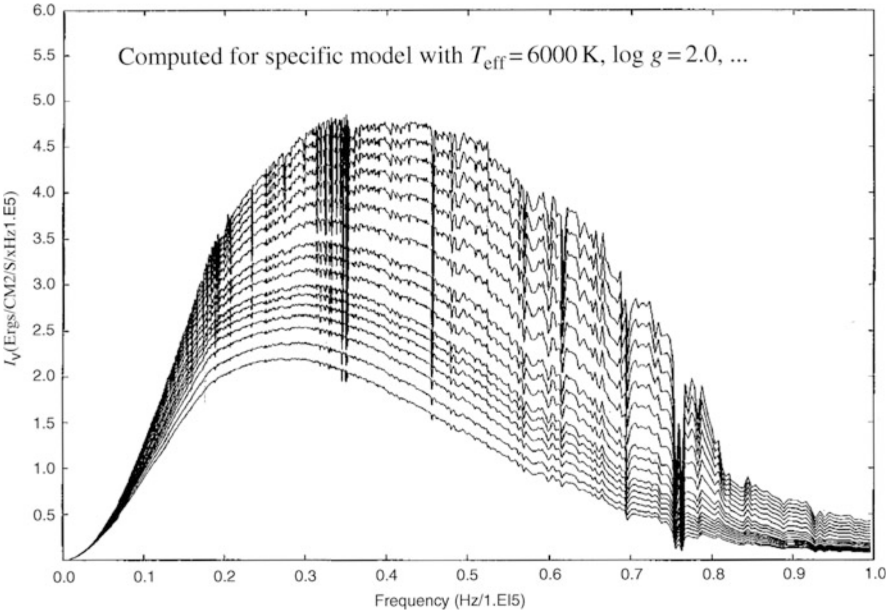


Fig. 4.7 Models of limb-darkening: computed stellar intensity for values of $\mu = \cos\theta = 1$, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.25, 0.2, 0.15, 0.1, 0.125, 0.1, 0.075, 0.05, 0.025, and 0.01, for images top to bottom, respectively (from Kurucz 1993)

This reflects the effect of the temperature reversal in the chromosphere and the possibility of contributions of atoms and ions from more than one level of the solar atmosphere.

From eclipsing binary star light curves, limb-darkening can also be measured on the disks of other stars. Equation (4.33) is most often used in this context, but even then, non-linear limb-darkening forms are sometimes found to be necessary to achieve satisfactory fitting of theoretical light curves to the branches of the light curve minima. Modern light curve analysis programs now have square-root and logarithmic as well as linear forms of the limb-darkening law (see Kallrath and Milone 2009, Sect. 3.2.4).

4.5.3 Granulation and the Convection Zone

The photosphere lies atop an extensive *convection zone*, in which rising parcels of gas expand and cool and descending parcels are compressed and become hotter. Observationally, the convection zone shows up as large cells of upwelling material about 10^3 km across known as *granules*. The upward speeds of the hot gas in these granules is ~500 km/s. This gas radiates as it nears the surface, cools, and sinks back down at the edges of the convective cells. These narrow cell boundaries appear darker by contrast. Larger regions of organized motions called *supergranules*, up to 3×10^4 km across, underlie the smaller cells. Although the upper velocities of these motions are smaller, they manage to concentrate magnetic fields in the lower chromosphere into networks, seen as heated regions in strong-line spectroheliograms.

Convection arises if the change in pressure with temperature through the stellar atmosphere is larger than would occur in a parcel of gas that rises *adiabatically*, i.e., without heat loss. Such an effect can occur when there is strongly absorbing material. At the core of the Sun, the temperature is expected to be $\sim 15 \times 10^6$ K, and the energy created through nuclear reactions is radiated away from the inner $0.25\mathfrak{R}_\odot$. At a distance of $\sim 0.71\mathfrak{R}_\odot$ from the center, the temperature has dropped to $\sim 2 \times 10^6$ K and the opacity of the matter to radiation begins to increase because of increased absorption. This occurs because highly ionized atoms begin to recapture electrons, and so are available to absorb photons. Whereas recapture results in photon emission, outgoing radiation is depleted as the absorbed radiation is reradiated in all directions. This increased opacity deep in the envelope of the Sun triggers convection, and this marks the start of the solar convection zone.

4.5.4 Other Photospheric Features

Above the outer edge of the convection zone, one of the principal sources of opacity in the solar photosphere is the absorption of radiation by the H^- ion, a hydrogen atom to which a second electron has been loosely attached. This was first suggested as a major source of the continuous opacity in the optical spectrum of the Sun by Rupert Wildt (1939).

Magnetic fields are observed on the face of the Sun; they probably arise from the convective motions of charged particles and become amplified close to the surface. The overall field of the Sun is relatively weak: $\sim 10^{-4}$ T. This can be compared to the value at the Earth's surface, $\sim 4 \times 10^{-5}$ T. From time to time, localized regions of strong magnetic fields appear at low to mid-latitudes. These fields inhibit convection and result in cooler and thus darker areas than the surrounding photosphere. We know these darker areas as sunspots.

The image of a sunspot group in the light of the core of the Hα line in Fig. 4.8 shows the detailed structure of the Sun in the vicinity of the spots, and demonstrates the strong influence of magnetic fields in shaping the structure. Active regions are given numerical designations; that shown is called AR8971. A flare can be seen in progress on the right; plages or faculae, brightenings around the spots, in discrete spectral lines and in white light, respectively, can also be seen. Such events are numerous near sunspot cycle maximum.

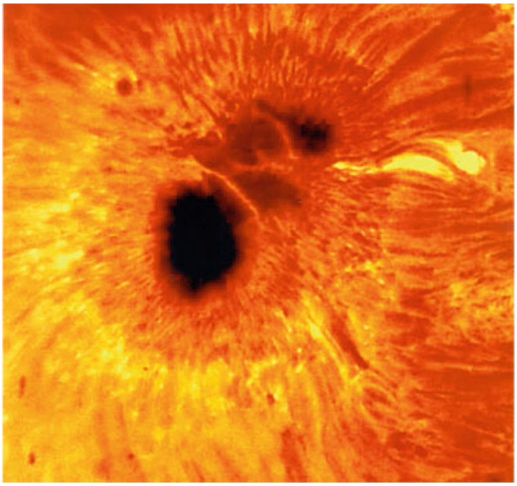


Fig. 4.8 An active sunspot group (AR 8971) as recorded in the Hα spectral line on April 27, 2000, with the McMath–Pierce Solar Telescope at the Kitt Peak National Observatory, Arizona. The area covered is 100 × 100 arc-secs. Courtesy Dr T. A. Clark. The McMath–Pierce Solar Telescope is operated for the National Science Foundation by the Association of Universities for Research in Astronomy as part of the National Solar Observatory

4.6 The Chromosphere

The temperature of the solar atmosphere decreases with increasing radius through the photosphere and reaches a minimum (~4,200 K) in the lower chromosphere. It then rises to an intermediate plateau before increasing rapidly in a *transition zone*, eventually reaching millions of degrees in the solar corona.

Different regions of the solar atmosphere can be studied by observations at different wavelengths within spectral line profiles. Strong lines saturate quickly in the line center which thus originates in the highest level of the solar atmosphere. In the "wings" of the line, the physical source depth is greater; the radiation comes from deeper layers of the atmosphere. Ultraviolet, x-ray, and radio regions of the spectrum are used to explore the outer solar atmosphere: the chromosphere, the transition region, and the corona. The chromosphere is the source of most of the ultraviolet radiation that impacts the upper atmosphere of the Earth.

The chromosphere extends upward from the photosphere for ~2,000 km. In this region, the gas density drops to ~10^{-4} that of the photosphere. Although it is 40× thicker, this layer has much weaker absorption and thus lower optical depth over most of the visible spectrum, so that it appears only ~10^{-4} as intense as the photosphere. The temperature varies only slightly with height, from a minimum of ~4,200 K to 25,000 K, over most of the chromosphere. Beginning ~2,000 km

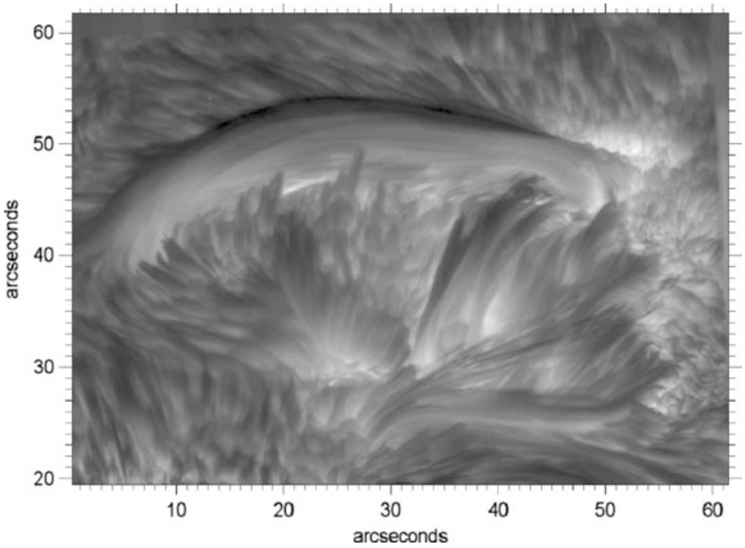


Fig. 4.9 The decaying active region AR 10812, as seen on October 4, 2005 in an image taken with the Swedish 1-m Solar Telescope by L. H. M. Rouppe van der Voort and M. J. van Noort, University of Oslo. A time series made of many carefully aligned images of this and another active region revealed movement of bright points with horizontal speeds of between 100 and 240 km/s, and somewhat slower features moving in loops. A bright surge is seen in the *top right part* of the image. Flame-like spicules are also visible. Courtesy of van Noort and Rouppe van der Voort (2006, Fig. 1), and reproduced by permission of the American Astronomical Society and the authors

above the photosphere, the density decreases rapidly by ~3 orders of magnitude and the temperature rises sharply to ~10^6 degrees.

Spectroheliograms taken in the cores of strong spectral lines, such as Hα, reveal the structure of the chromosphere and clearly show the presence of a *chromospheric network*, a region of concentrated magnetic fields that contain the enigmatic *spicules*. The fine structure in an active region (AR10812) in Hα can be seen in Fig. 4.9, from Fig. 1b of van Noort and Ruppe van der Voort (2006). Plages, pores, surges, and spicules are visible.

The chromosphere (and at very short wavelengths, the lower corona) is revealed also in the spectral lines of the far UV, where the opacity is also high. At wavelengths below about 1600 Å (160 nm), the solar spectrum changes from an absorption to an emission spectrum, indicating the absence of cooler radiation layers above the emitting region. Figure 4.10 displays the spectrum of the Sun in the "rocket ultraviolet," so-called because the Earth's atmosphere will not pass radiation at wavelengths shorter than ~320 nm. The echelle spectrograph used to obtain this spectrum was carried aloft on an Aerobee rocket to an altitude of ~100 km above the Earth's surface. A spectral resolution element as fine as 0.020 Å was achieved (Tousey et al. 1974).

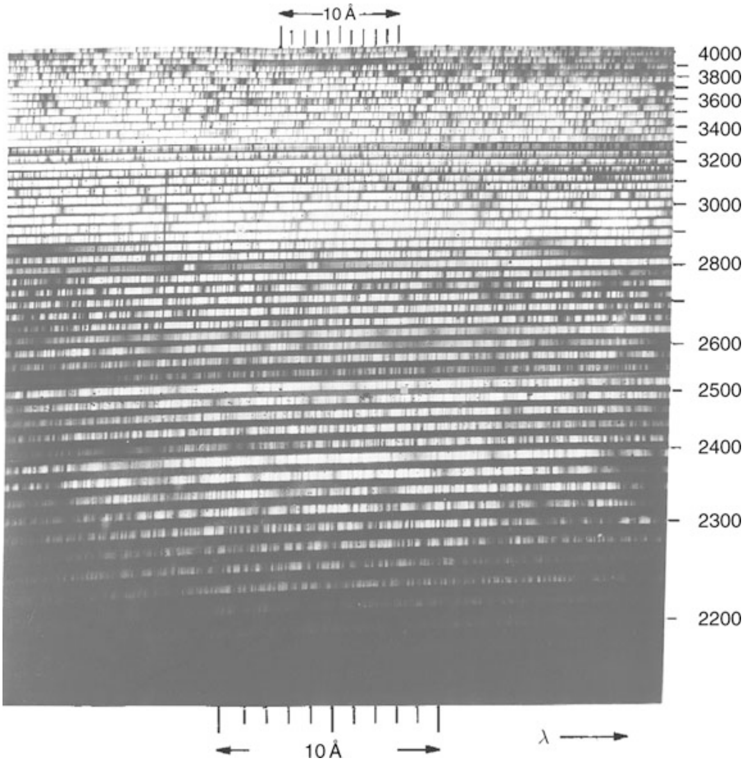


Fig. 4.10 The ultraviolet spectrum of the center third of the solar disk, obtained with an echelle spectrograph on an Aerobee rocket in 1961. The prominent bifurcated emission features superimposed on broad absorption lines near 2800 Å = 280 nm are the h and k lines of MG II. Courtesy Dr. R. Tousey, US Naval Research Laboratory

Radiation from higher levels is also seen at visual wavelengths to a certain extent, particularly in the cores of the strong absorption lines of the Balmer series of hydrogen, Ca II (H and K), and in Mg II (h and k) lines, where the emission is enhanced. Total solar eclipse observations are also very important because they provide good signal-to-noise measurements when the lower layers are masked out by the lunar disk. They can also provide high-resolution measurements, and those made from space are undisturbed by terrestrial atmospheric seeing effects.

From the SkyLab mission ~1974, from which the solar ultraviolet limb-darkening was studied, much was learned about the structure of active regions throughout the solar atmosphere. Figure 4.11 shows the image of the Sun in the 304 Å = 30.4 nm line of helium II, the equivalent of the Lyman α line (of hydrogen) for ionized helium. The adjacent image is of a transition involving a highly ionized atom of iron. The solar disk is actually seen only in silhouette, outlined by active regions on the limb, and through the thin emission from the chromosphere and lower corona.

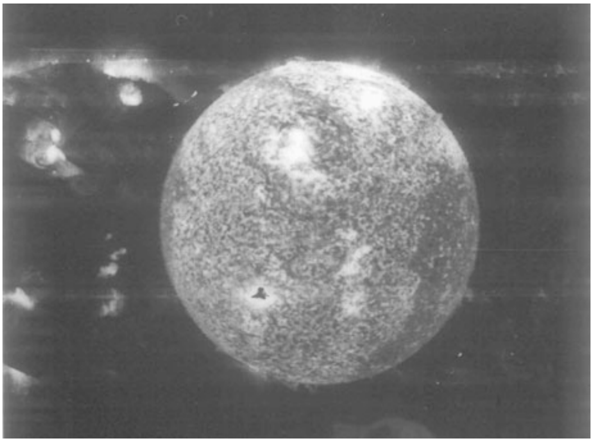


Fig. 4.11 Far ultraviolet images of the Sun from NRL's slitless spectrograph on board SkyLab. The *central image* is in the light of the helium II line at $\lambda = 30.4$ nm. The *dark region* to the *lower right* is a *coronal hole*, where the particles are allowed to stream away from the Sun as the solar wind (see Milone and Wilson 2014, Chap. 11.3), resulting in a reduction in photographic density. In the *lower left*, an intense flare is so bright at its center that it has *decreased* the photographic density (a condition known as *solarization*). A second image, to the *left*, is the appearance of the Sun in the light of a transition of a highly ionized species of iron. Note that in the iron-line image, the disc of the Sun is seen only in silhouette against the glow from active regions behind the limb. Courtesy Dr. R. Tousey, US Naval Research Laboratory

4.7 The Corona

This physically thick but extremely tenuous collection of gas, ionized particles, and dust has been likened to a thin flame. We are able to see the photosphere and even the chromosphere through it. The corona extends several solar radii out into the solar system, yet the intensity is $\sim 10^{-6}$ that of the photosphere. It is the extremely low density ($\sim 10^{5}$particles/cm^{3}) that makes this layer so transparent.

The corona has three main components: the *K-corona*, the *F-corona*, and the *E-corona*.

The K-corona is the light of the photosphere scattered by electrons in the corona. The name derives from the German *kontinuierlich*, continuous. This contribution is most important between 1 and $2.3\mathfrak{R}_{\odot}$, where it dominates the coronal light. The K-corona varies in appearance and polarization across the solar cycle. The polarization can be as high as 70 % in the inner middle corona.

The F-corona is the contribution of scattered photospheric light by circum-solar dust grains. This scattering includes the absorption or Fraunhofer lines (so-called because Fraunhofer first detected these absorption lines in the Sun). The F-corona contributes to the *zodiacal light*, best seen after sunset in the Spring, or before sunrise in the Fall, when the ecliptic nearest the Sun stands most vertically above the horizon. The dust absorbs as well as scatters sunlight and reradiates in the infrared.

The E-corona is the source of emission lines from highly ionized atoms and is observed from all regions of the corona. Edlén (1942) identified many of these lines following a suggestion by Grotrian (1939). The very high temperatures produce high levels of ionization and these temperatures along with the very low densities create a situation where even highly unlikely transitions can occur, giving rise to *forbidden line* emission. Normal transitions occur over intervals of 10^{-8} s. Normally particle collisions are frequent enough in gases that only these likely transitions can occur, giving rise to spectral lines of permitted transitions. Because collisions are relatively rare in the corona, radiative transitions from *metastable* levels, requiring intervals of the order of seconds, can occur.

The evidence for high temperatures in the corona comes from observations of all three coronal components. Fraunhofer lines are washed out in the K-corona, implying very high speeds of the electrons (Grotrian 1931). The F-corona, scattered by much heavier dust particles, preserves the absorption lines better, but with some smearing. The emission lines of the E-corona arise from species of ions which can only appear in a very high temperature environment.

4.8 The Sunspot Cycle and Solar Magnetic Cycle

A few features of the active Sun were mentioned in Sect. 4.5.4. The frequency and level of activity varies, however. In this section we discuss that variation.

The average period of variation in the numbers of, and total area occupied by, sunspots is about 11 years (Fig. 4.12). Beginning at solar minimum, the first spots of each *sunspot cycle* form at ~30°N and S latitude. As the cycle progresses, the mean latitude in each hemisphere decreases toward the equator while the number of spots increases to solar maximum and decreases again toward the next solar minimum. The last spots of the cycle form within ±~10° of the equator. This variation in number and latitude of sunspots results in a pattern similar to butterfly wings in the upper part of Fig. 4.12, giving the figure its name, eponymously after Edward W. Maunder (1904), the *Maunder butterfly diagram.*

The solar magnetic dynamo is located in the convection zone in the outer part of the Sun (Sect. 4.5.3). The Sun rotates differentially, with a rotation period varying from ~25 days at the equator to ~35 days at high latitudes. As a result, the magnetic field lines below the surface are carried around the Sun progressively faster at lower latitudes, stretching the field lines in a direction approximately parallel to the solar equator. This stretching intensifies the magnetic field, and sunspots form in the patterns described above when buoyancy carries loops of magnetic field out through the photosphere.

Sunspots thus frequently occur in pairs of opposite magnetic polarity, with one of the pair leading the other as the Sun rotates. In a given solar cycle, if the leading spot in each pair in the northern hemisphere has north magnetic polarity, then the leading spot in the southern hemisphere will have south magnetic polarity. The magnetic field of the Sun reverses every sunspot cycle, with the result that the

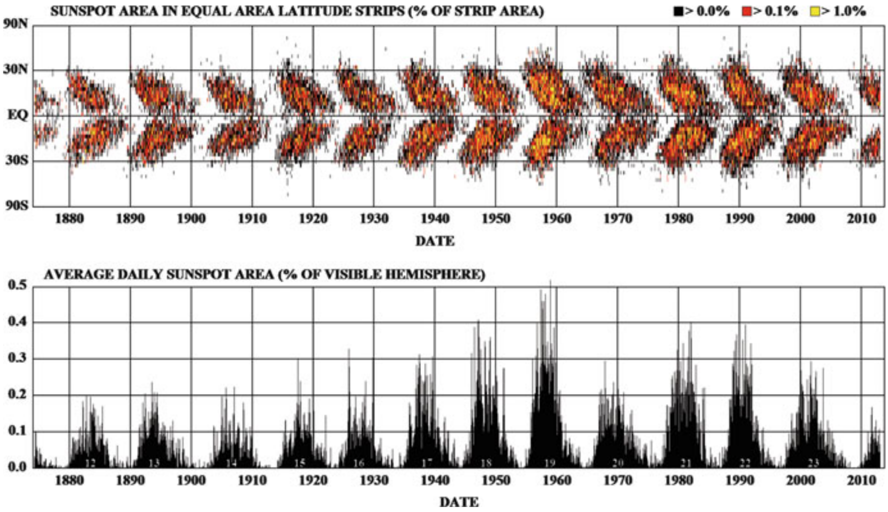


Fig. 4.12 Daily sunspot area averaged over individual solar rotations. *Upper figure*: Maunder butterfly diagram showing the area occupied by sunspots in equal-area latitude strips (fraction of strip area) as a function of latitude and time. *Lower figure*: Total sunspot area as a function of time. Credit: Hathaway/NASA

polarities of the sunspot pairs reverse also. The resulting solar magnetic cycle has a mean period equal to two sunspot cycles, or ~22 years.

The magnetic field suppresses convection in the sunspots, resulting in lower temperatures than in the surrounding photosphere, and causing them to appear dark in comparison. However, convective and bulk motions of gas from around the spots increase with solar activity. This solar activity leads to variable line profiles in spectral lines, especially in emission features at the centers of CaII H and H lines at 393.3 and 396.8 nm respectively. Such variations in the spectra of other stars would thus indicate equivalent stellar activity.

Solar flares occur in the regions of strong magnetic fields near sunspot groups. Magnetic fields store energy, with the energy density, u, increasing with increasing magnetic field strength, B: $u = B^2/(2\mu_0)$, where $\mu_0 = 4\pi \times 10^{-7}$ N/A^2 is the permeability of free space, and the units N and A are newtons and amperes, respectively. When these fields become twisted by material motion around the sunspots, *magnetic reconnection* can take place as illustrated schematically in Fig. 4.13. This magnetic realignment can release several times 10^{25} J of energy in times of only a few minutes to a few tens of minutes, creating the solar flare.

The activity level affects also the appearance of the corona, which tends to become the largest and most spherically symmetric at solar maximum. The corona may contain material trapped in closed magnetic fields extending above bipolar spots. These are *prominences*, with ~100× greater density but only 1/100 the temperature of the surrounding corona. They are seen as bright arcs, i.e., in emission above the solar limb; but when projected against the disk they are seen

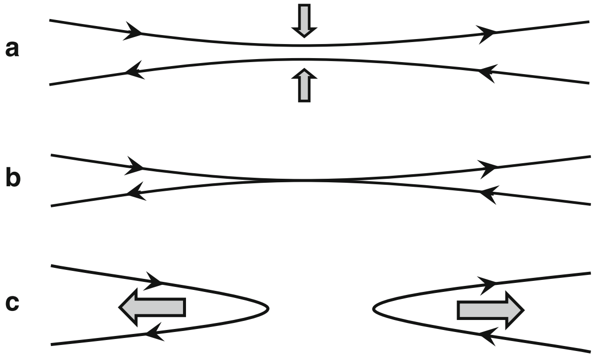


Fig. 4.13 Magnetic reconnection. If two magnetic field lines, oriented in opposite directions, approach each other as in (**a**), they can merge (**b**) and then separate again (**c**). Reconnection converts potential energy stored in the magnetic field in (**a**) into kinetic energy of the plasma as the field lines separate in (**c**)

in absorption, and appear as dark *filaments*. Unlike flares, *quiescent prominences* may be visible for days while *eruptive prominences* can expand outward from the Sun over times of several hours.

Coronal mass ejections (CMEs) are another high-energy phenomenon of the active Sun, often seen as a loop of material expanding outward through the corona and into interplanetary space on time scales of several minutes to several hours (see Webb and Howard (2012) for a review of observations). Their typical frequency of occurrence is ~5/day at solar maximum and ~1/day at solar minimum. The largest can carry $>10^{13}$ kg of plasma at speeds >2,500 km/s, with a total mechanical energy requirement (to escape the Sun's gravitational field at the given speed) of $>10^{25}$ J. Many are associated with solar flares in active regions, although they can also be associated with eruptive prominences located away from active regions. There are, however, indications that eruptive prominences are associated with solar flares, so the two apparent categories of CMEs may in fact be end members of a continuum of events with a common source, rather than being physically separate phenomena. It should also be noted that not all CMEs are associated with flares, and certainly not all flares produce mass ejection; also, the energy involved in a CME is much greater than that in a flare. The common association of CMEs and solar flares therefore does not mean that one causes the other; more probably, they are separate results of a single magnetically-driven event.

Solar activity, such as flares or eruptive prominences, is often accompanied by bursts of radio emission in the kHz to GHz range. The intensity and frequency (number of bursts per unit time) vary over a solar cycle, peaking near solar maximum. Five types of radio bursts are recognized:

Type I: Frequency range ~$80 < f < 200$ MHz; narrow bandwidth ($\Delta f \sim$ 5–30 MHz). Individual bursts last typically ~1 s, but occur in "burst storms" that can last from hours to days. They are 100% circularly polarized, indicating an origin in strong magnetic fields, and their source regions appear to lie in the loops of magnetic field above sunspots and sunspot groups. The specific emission mechanism, however, is not yet clear. For a brief review, see Ramesh and Shanmugha Sundaram (2000).

Type II: Frequency range ~40 $< f <$ 150 MHz; narrow-bandwidth ($\Delta f/f$ often 0.2–0.4). Each burst typically drifts slowly from higher to lower frequency over times from ~1 to 30 min, and is often accompanied by a harmonic at a frequency ratio of ~2:1. The bursts originate in outwardly-travelling shock waves in the solar corona that are produced by solar flares and/or CMEs (Mann et al. 1995). The decreasing electron density as the shock propagates outward through the corona produces the decreasing frequency.

Type III: Frequency range from $f \approx$ 1 GHz at the bottom of the corona to 30 kHz at $\approx$ 1 au. Each burst typically shows a fast (1–3 s) drift toward lower frequencies, and bursts often occur in groups lasting up to a few minutes, or "storms" of up to a few hours. The bursts originate in a three-step process (Ginzburg and Zheleznyakov 1958; Thejappa et al. 2012; Saint-Hilaire et al. 2013): (1) A particle-acceleration event such as a flare or CME creates an electron beam travelling outward along magnetic field lines into the heliosphere. (2) The beam excites Langmuir waves (oscillations in electron density) in the solar plasma in a very narrow band around the plasma frequency, $f_{pe} = 9n_e^{1/2}$, where n_e is the electron density in m^{-3}. (3) Nonlinear plasma processes convert the Langmuir waves into electromagnetic waves at f_{pe} and $2f_{pe}$. The frequency decreases in response to the decrease in n_e as the electron beam propagates outward.

Type IV: Characterized by persistent smooth broadband continuum emission that can last from minutes to days and extend over frequencies from ~10 MHz to ~10 GHz (Weiss 1963); often preceded by a Type II disturbance, although not necessarily at the same position. Often, fine burst structure is superimposed on the continuum emission. They are almost always associated with solar flares and CMEs, and are of particular interest because they are often followed, after a suitable delay due to travel time, by the arrival of solar protons at the Earth and the occurrence of geomagnetic disturbances. At least two subtypes are recognized: "stationary" Type IV bursts that can last from hours to days, and "moving" Type IV bursts that usually last from 0.5 to 2 h and show rapid motion outward through the solar corona. Moving Type IV bursts have been shown to arise from either plasma or synchrotron emission, the former triggered by shock wave propagation or by beams of accelerated electrons, and the latter from mildly relativistic electrons gyrating in magnetic fields. They have also been observed from different components of CMEs: advancing fronts, expanding loops, and isolated sources. See Tun and Vourlidas (2013) for a brief review and references.

Type V: Short-duration, smooth continuum events that are often observed at the end of a Type III burst (and never otherwise).

Historically the sunspot cycle has proven variable both in cycle length, and, more spectacularly, in strength, i.e., in the sunspot number. In some cycles, the sunspot maximum is greatly diminished in the sense that few if any sunspots are seen; this diminution may extend over several cycles. A number of such activity minima have been noted, the most prominent being the *Maunder minimum* of 1645–1716. Eddy (1976) in an excellent summary notes that the sum of all spots recorded in this entire 70-year interval was less than that seen during a "single normal year of active conditions." In the current sunspot cycle, activity indicators suggest that another such epoch of diminished activity may be starting (Pasachoff and MacRobert 2011).

For a fuller examination of the solar cycle as well as the quiet Sun, we can still recommend Gibson (1973) and, for solar phenomena throughout the solar cycle, Brandt (1970). Golub and Pasachoff (2010) do justice to the rich history of coronal studies, and Bhatnagar and Livingston (2005) cover the development of solar physics as well as the phenomena as viewed by cultures throughout history, comprehensively.

4.9 Line Absorption and Emission

The discussion in Sect. 4.5.1 can be expanded to show how absorption and emission lines can be understood. Beginning with (4.20), and writing $\kappa_\lambda = k_\lambda \rho$, for convenience, we arrive at this general form of the *transfer equation*:

$$\mathrm{d}I_\lambda/\mathrm{d}r = -I_\lambda \kappa_\lambda + \varepsilon_\lambda \tag{4.34}$$

where ϵ_λ is the emissivity per unit length. After dividing through by κ_λ, and substituting $\mathrm{d}\tau_\lambda = \kappa_\lambda\, \mathrm{d}r$, from (4.17), where τ_λ is the optical depth, we get:

$$\mathrm{d}I_\lambda/\mathrm{d}\tau_\lambda = S_\lambda - I_\lambda \tag{4.35}$$

where S_λ is the *source function*, $\mathrm{S}_\lambda \equiv \epsilon_\lambda/\kappa_\lambda$ (W m^{-2} sr^{-1} Hz^{-1}).

Re-arranging (4.35),

$$\frac{dI_\lambda}{I_\lambda - S_\lambda} = -d\tau_\lambda \tag{4.36}$$

Integrating from $\mathrm{I}_{0,\lambda}$ at $\tau_\lambda = 0$ to I_λ at τ_λ with the assumption that S_λ is constant gives

$$\ln \frac{I_\lambda - S_\lambda}{I_{0,\lambda} - S_\lambda} = -\tau_\lambda \tag{4.37}$$

Taking the inverse log,

$$I_\lambda - S_\lambda = (I_{0,\lambda} - S_\lambda)\exp(-\tau_\lambda) \tag{4.38}$$

and solving for I$_\lambda$,

$$I_\lambda = I_0\exp(-\tau_\lambda) + S_\lambda[1 - \exp(-\tau_\lambda)] \tag{4.39}$$

where $I_{0,\lambda}$ is the intensity of the beam emerging from the deeper atmosphere into the detectable region defined by $0 \to \tau_\lambda$, and S$_\lambda$ is the emission added in this region.

Now we can consider two special cases:

1. When $I_0 = 0$, the emerging intensity is merely:

$$I_\lambda = S_\lambda[1 - \exp(-\tau_\lambda)] \tag{4.40}$$

There are two further possibilities within this first case:

A. If the parcel is *optically thin*, so that $\tau_\lambda \ll 1$, then $\exp(-\tau_\lambda) \approx 1 - \tau_\lambda$, and

$$I_\lambda \approx S_\lambda[1 - (1 - \tau_\lambda)] = \tau_\lambda S_\lambda \tag{4.41}$$

B. If the gas is *optically thick*, so that $\tau_\lambda \gg 1$, then $\exp(-\tau_\lambda) \to 0$, and

$$I_\lambda \to S_\lambda \tag{4.42}$$

In local thermodynamic equilibrium (LTE), such as occurs at large optical depth inside a star, $I_\lambda = B_\lambda(T)$, where T is the temperature of the material (see Sect. 4.4.3). Thus, in LTE, $S_\lambda = B_\lambda(T)$ also.

2. When $I_0 \neq 0$, for an optically thin gas parcel,

$$I_\lambda \approx I_0(1 - \tau_\lambda) + \tau_\lambda S_\lambda = I_0 + \tau_\lambda(S_\lambda - I_0) \tag{4.43}$$

When $I_0 > S_\lambda$, the second term on the right-hand side (RHS) is negative so the RHS is, in fact, less than I_0, and therefore $I_\lambda < I_0$. An absorption line results.

However, when $I_0 < S_\lambda$, we have added more emission than entered this parcel, $I_\lambda > I_0$, and an emission line is seen at this point in the spectrum.

4.10 Helioseismology

Many stars are variable in light and in mean radial velocity. Some stars are "geometric" variables, changing brightness because of eclipses by companions, or, by virtue of rotation of the star, revealing irregularities of shape (as in close binary systems) or surface brightness (starspots, faculae, etc.). Others vary because

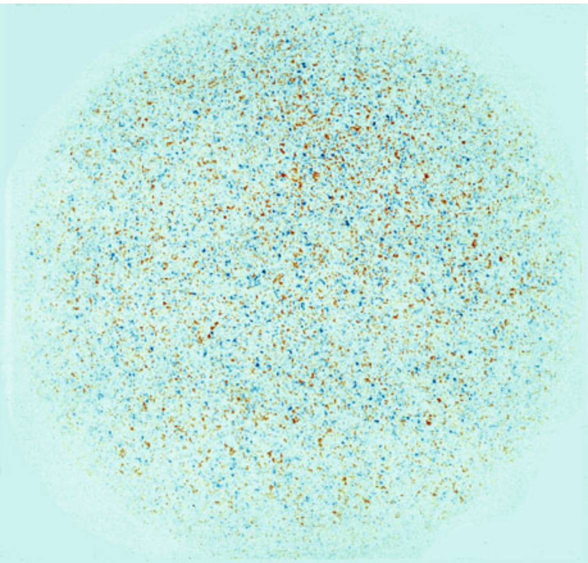


Fig. 4.14 Illustration reprinted with permission from Harvey (1995, Fig. 1). The intensity signature of hot rising (*red*) and falling (*blue*) gas from two images obtained 2 min apart at an observatory at the South Pole in January, 1991. Copyright 1995, American Institute of Physics

of large-scale motions of their atmospheres, as RR Lyrae, delta Scuti, or Cepheid variables.

The Sun has long been known to be variable, by virtue of its sunspots, on both short (rotation period) and long (the solar cycle) time scales in virtually all regions of the electromagnetic spectrum. Visually, this variability is at a very low level: the area occupied by sunspots very rarely exceeds 1% of the solar disk. Over the past few decades, the Sun has been shown to pulsate also (Leighton 1961; Leighton et al. 1962). These pulsations have periods of about 5 min and very small amplitude, but they show up particularly well in some spectral features. These solar oscillations reveal the interior structure of the Sun. In Harvey's (1995) words, the oscillation spectrum is "rich and crowded." The process of retrieving the information is complicated and beyond the scope of our discussions. The following description of these waves and their significance can be found in more detail in Harvey (1995). The waves show up in two ways:

1. *Doppler shifts in the spectral lines*. The velocity variations are very small, ~ 0.1 m/s, so that very high spectral resolution is required. Only in the Sun does one have the flux to observe such a variation with precision, because one needs high signal/noise capability and very high spectral dispersion. The techniques needed to attain this are beyond the current topic of discussion, but often involve Fourier Transform Spectroscopy techniques. Figure 4.14 illustrates the effect on the disk of the Sun, and Fig. 4.15 summarizes the results of the time-period analysis of a 3.5-day run of nearly continuous observations from Antarctica (Grec et al. 1980).
2. *Surface brightness variations*. Because the wave motion may cause material to oscillate in position, there are brightness variations accompanying the

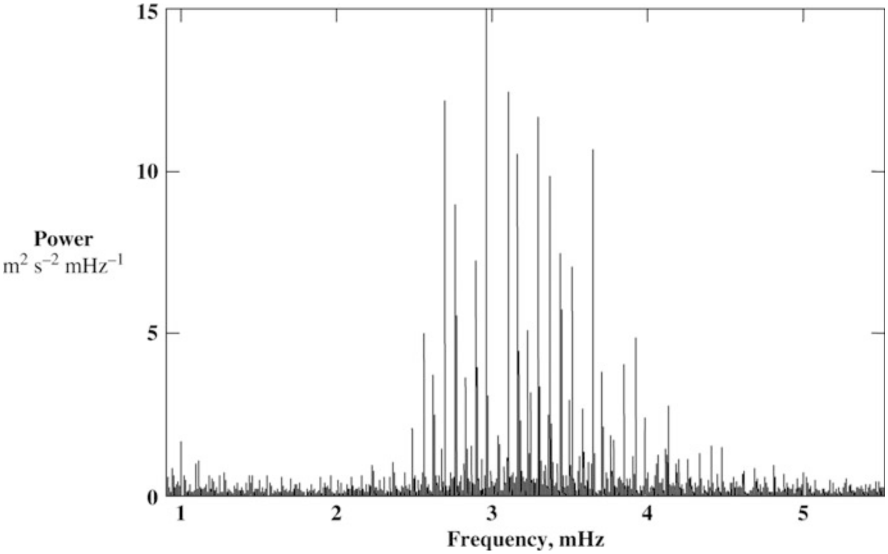


Fig. 4.15 A Doppler spectrum from the whole solar disk taken over 6 days of nearly continuous coverage from an Antarctic observatory, showing various frequencies (inverse of periods) of gas motion. One of the strongest peaks is at ~0.0033 Hz, (a period of about 5 min), representing radial mode $n = 23$ and spherical harmonic $\ell = 0$. Adapted from Grec et al. (1983)

oscillation. The variation is again tiny: a few *micro*magnitudes. Stellar astronomers are doing well if they can establish light parameters to ±0.001 magnitude, although large telescopes, long integration times, and improved detectors have enabled this level of precision, if not accuracy, to be attained more easily in recent years. Such very high precision is very difficult to obtain, and only the very high flux levels of sunlight provide sufficient precision of measurement to permit it, but exacting means to deal with stray light and a host of other issues must be applied.

There are four basic types of nonmagnetic waves of general interest in the Sun. These are: acoustic waves, surface gravity waves, internal gravity waves, and Rossby waves. Only the first two produce 5-min oscillations and are unambiguously observed. Resonant cavities within the Sun "organize" these waves into standing wave patterns, which are called *p*-, *f*-, *g*-, and *r*-modes, respectively. Over ten million *p*- and *f*-modes alone have been shown to exist. The global *p*-modes are seen in Fig. 4.15. The origin of a wave or series of waves may be the displacement of matter and a restoring force acting on the gas. In the case of acoustic waves this restoring force is the gas pressure, hence *p*-modes. In the case of internal gravity waves (*g*-modes), it is the buoyancy or vertical pressure forces. Similarly the latter serve as restoring forces for the surface gravity waves (*f*-modes, for *fundamental*) that propagate at the interface between the photosphere and the chromosphere but, if low angular harmonics are excited, can penetrate far into the sun.

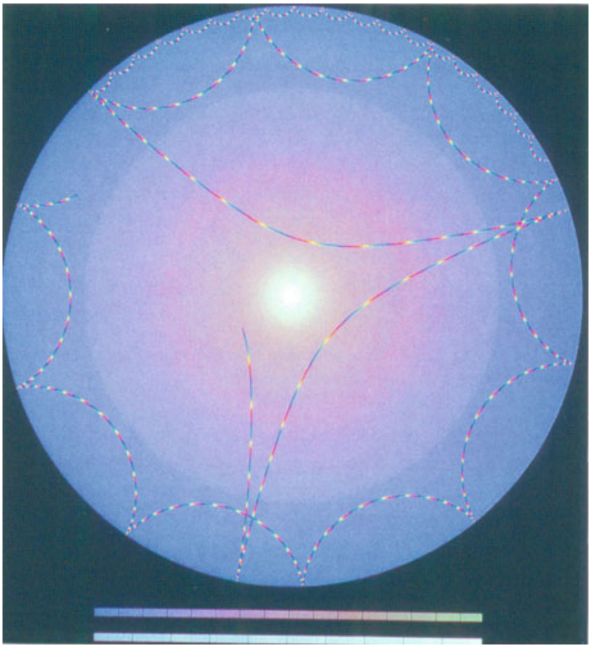


Fig. 4.16 A cutaway model of the Sun, showing refracted acoustic rays (*dashed curved lines*) and the temperature (*color*, *increasing from left to right along the legend bar*) and energy production (intensity) through the solar interior. Ray calculations were done by Dr. D. D'Silva. Illustration reprinted with permission from Harvey (1995, Fig. 2). Copyright 1995, American Institute of Physics

The region which traps acoustic waves in the Sun is defined by two regimes: the lower boundary where the sound speed increases[7] and the upper boundary, near the surface, where the density decreases sharply. Figure 4.16 illustrates this. The number of radial nodes, n, characterizes the standing wave. The radial nodes are depicted in Fig. 4.17, where the radial nodes run from 0 (lower right) to 40 (upper left).

The standing waves on the Sun's circumference can be seen on great circles such as longitude circles and on small circles, such as latitude circles. These are characterized by the spherical harmonic, ℓ, and by the azimuthal order, *m*. The depth of the 'cavity' is revealed by the quantity ℓ/f, where f is the cyclic frequency, in the sense that the smaller this ratio, the deeper the cavity. The expected wavelengths, ℓ_n, and frequencies, f_ℓ of acoustic waves are determined by the radial distance, r, of the cavity bottom from the Sun's center, and by ℓ, the nodal lines (circles):

$$\lambda_n = (2\pi r)/\sqrt{\ell(\ell+1)} \tag{4.44}$$

or

[7] The result of this is to cause downward-propagating waves to bend back toward the surface.

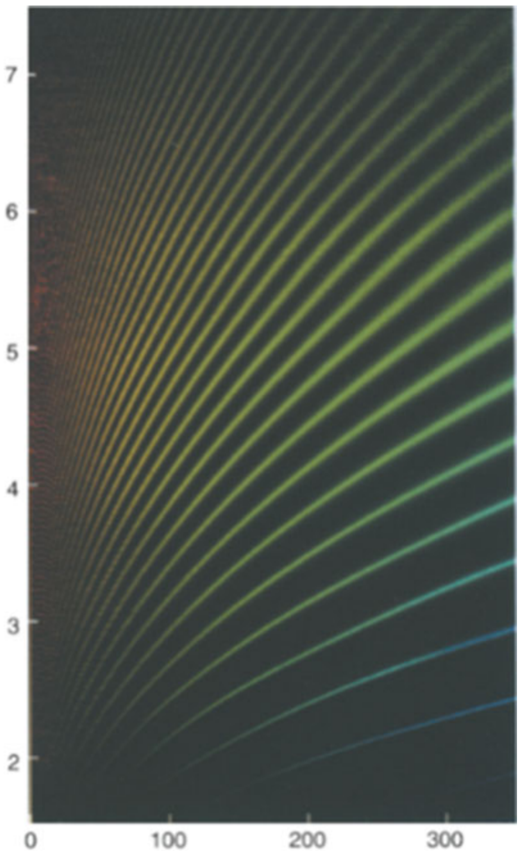


Fig. 4.17 The power spectrum of solar oscillations in the form of standing acoustic wave patterns. The frequency in mHz is plotted against the spherical harmonic degree, ℓ. *Color* represents depths, with *blue* indicating shallow and *red* indicating deep, penetrating modes. Image supplied by John W. Harvey. Illustration reprinted with permission from Harvey (1995, Fig. 3). Copyright 1995, American Institute of Physics

$$f_{\ell} = (2\pi)/t_{\lambda n} = (2\pi v_s)/\left[(2\pi r)/\sqrt{\ell(\ell+1)}\right] \tag{4.45}$$

$$= (\gamma P/\rho)^{1/2}[\ell(\ell+1)]^{1/2}/r \tag{4.46}$$

where $t_{\lambda n} = \lambda_n/v_s$ is the period, v_s is the speed of sound, γ is the ratio of specific heats, P is the pressure, and ρ the density (see Milone and Wilson 2014, Chap. 10.2 for definitions and discussion). As

$$v_s \propto \sqrt{T}, \tag{4.47}$$

and T increases inward (from the temperature minimum), both v_s and f_λ increase with depth.

Internal gravity waves are thought to be important only below the convection zone and are extremely weak at the surface, with velocity amplitudes expected to be

<0.3 mm/s at ~100 μHz frequency (~3 h period). The *buoyant frequency*, sometimes known as the *Brunt–Väisälä frequency* for this mode of oscillations is:

$$f = \sqrt{g[(1/\gamma P)\partial P/\partial r - (1/\rho)\partial \rho/\partial r]} \tag{4.48}$$

where g is the gravitational acceleration at r.

To our knowledge, there has been no undisputed detection of g-mode solar oscillations to the present time, but Wolff (2002) suggests that maxima in a 50-year record of 10.7 cm solar radio flux data provide evidence of solar core rotation (at ~430 nHz) and g-mode beat frequencies (below ~60 nHz). Wolff's (2009) mixed shell model for the inner 25% of the solar radius makes it much less difficult to excite g modes. The p-modes that are seen already tell a great deal about the Sun's structure; ray tracing is another informative technique, just as the arrival times of seismic waves at various locations on Earth's surface indicate the structure of the Earth's interior. A downward-propagating acoustic wave is refracted and traverses a distance over a particular time interval that depends on the conditions along the path. The time intervals and locations on the Sun's surface are determined not by discrete "quakes", as on Earth, but on correlations among the waves seen in ringed areas (annuli) around particular points. Figure 4.16 illustrates three types of refracted wave paths. As noted above, the sound speed increases inward, so ingoing (non-vertical) waves curve upward. Following the predictions of Ulrich (1970), Leibacher and Stein (1971), and of Wolff (1972), who first noted the global character of the oscillations, the first clear detection of predicted p-mode signatures was by Deubner (1975). The oscillations require the base of the convection zone, once thought to be 0.80–0.85$\mathfrak{R}_\odot$, to be substantially deeper (Gough 1977). The best value at present for the base of the convection zone is 0.713 ± 0.003$\mathfrak{R}_\odot$. Another result of oscillation studies is that the speed of sound between 0.3 and 0.5$\mathfrak{R}_\odot$ was found to be larger than expected from solar models. The opacity assumed for the matter in this region had been too low, and subsequent work at Los Alamos confirmed this suspicion.

This concludes our brief examination of the Sun and of solar physics. Many of the ideas discussed here are applicable to other areas of solar system studies. We next treat the basic properties of those planets nearest to the Sun, the terrestrial planets.

Challenges

[4.1] From the solar parallax, the mean measured semi-diameter, and the solar constant, calculate (a) the radius of the Sun in linear measure and (b) the luminosity of the Sun.

[4.2] Beginning with the results of Q [4.1], compute (a) the effective temperature of the Sun and (b) the mean intensity of the Sun.

[4.3] Find the Sun's (a) apparent bolometric magnitude, m_{bol}, and its (b) absolute visual magnitude, M_{v}. (c) How bright would the Sun appear if it were at the mean distance of the Andromeda galaxy (M31), 0.75 Mpc (1 Mpc $= 10^6$ pc) away?

[4.4] Determine the wavelengths of peak emission for the Planck functions of Fig. 4.3, and compare the total wavelength-integrated flux emitted by these two black bodies. Finally, if the cooler of these two objects had twice the luminosity of the hotter, compute the ratio of their diameters.

[4.5] Determine the linear sizes of the sunspots, flare, and smallest features that can be discerned in Fig. 4.8.

[4.6] If you were observing the Sun from a nearby star system, and had a large telescope and instruments of high-precision and accuracy, what photometric and spectroscopic effects would you expect to see over the 22-year solar activity cycle?

References

Ahmad, Q.R., et al.: Measurement of the rate of $\nu_e + d \rightarrow p + e^-$ interactions produced by ^{8}B Solar Neutrinos at the Sudbury Neutrino Observatory, *Phys. Rev. Lett.* **87**(7), 071301 (2001)

Allen, C.W.: *Astrophysical Quantities*, 3rd edn. Athlone Press, London (1973)

Bahcall, J.N., Serenelli, A.M., Basu, S.: New solar opacities, abundances, helioseismology, and neutrino fluxes. *Astrophys. J.* **621**, L85–L88 (2005)

Beatty, J.K., Chaikin, A. (eds.): *The New Solar System*, 3rd edn. Sky & Telescope, University Press, Cambridge, MA, Cambridge, UK (1990)

Bhatnagar, A., Livingston, W.: *Fundamentals of Solar Astronomy*. World Scientific Publishing Co. Pte. Ltd, Singapore (2005)

Brandt, J.C.: *Introduction to the Solar Wind*. W.H. Freeman and Company, San Francisco (1970)

Brown, T.M., Christensen-Dalsgaard, J.: Accurate determination of the solar photospheric radius. *Astrophys. J.* **500**, L195–L198 (1998)

Cox, A.N. (ed.): *Allen's Astrophysical Quantities*, 4th edn. Springer-Verlag, New York (2000)

Deubner, F.-L.: Observations of low wavenumber nonradial eigenmodes of the sun. *Astron. Astrophys.* **44**, 371–375 (1975)

Dufay, J., tr. Gingerich, O.: *Introduction to Astrophysics: The Stars*. Dover, New York (1964). (tr. of *Introduction à l'astrophysiques: les ètoiles*, 1961, George Newnes)

Eddy, J.A.: The Maunder Minimum. *Science* **192**, 1189–1202 (1976)

Edlén, B.: Die Deutung der Emissionslinien im Spektrum der Sonnenkorona. *Zeitschrift für Astrophysik* **22**, 30–64 (1942)

Gibson, E.G.: *The Quiet Sun*. NASA SP-303 (1973)

Ginzburg, V.L., Zheleznyakov, V.V.: On the possible mechanisms of sporadic solar radio emission (radiation in an isotropic plasma). *Sov Astron* **2**, 653 (1958)

Golub, L., Pasachoff, J.M.: The Solar Corona, 2nd edn. Cambridge University Press, Cambridge (2010)

Gough, D.O.: In: Bonnet, R.M., Delache, P. (eds.) The Energy Balance and Hydrodynamics of the Solar Chromosphere and Corona, pp. 3–36. G. De Bussac Clermont-Ferrand (1977)

Gough, D.O.: How oblate is the Sun? *Science* **337**, 1611–1612 (2012)

Grec, G., Fossat, E., Pomerantz, M.: Solar oscillations: full disk observations from the geographic South Pole. *Nature* **288**, 541–544 (1980)

Grec, G., Fossat, E., Pomerantz, M.: Full-disk observations of solar oscillations from the geographic South Pole: latest results. *Solar Phys.* **82**, 55–66 (1983)

Grotrian, W.R.W.: Zur Frage der Deutung der Linien im Spektrum der Sonnenkorona. *Naturwissenschaften* **27**, 214 (1939)

Grotrian, W.R.W.: Ergebnisse der Potsdamer Expedition zur Beobachtung der Sonnenfinsternis am 9 Mai 1929 in Takengon (Nordsumatra) 6. Über die Intensit*ä*tsverteilung des kontinuierlichen Spektrums der inneren Korona, *Zeitschrift für Astrophysik*, 3, pp. 199–226 (1931)

Harvey, J.: Helioseismology. *Phys. Today* **48**(October), 32–38 (1995)

Kallrath, J., Milone, E.F.: *Eclipsing Binary Stars: Modeling and Analysis*, 2nd edn. Springer Verlag-Publishers, New York (2009)

Kjeldseth Moe, O., Milone, E.F.: Limb darkening 1945A to 3245A for the quiet Sun from SkyLab data. *Astrophys. J.* **225**, 301–314 (1978)

Kuhn, J.R., Bush, R., Emilio, M., Scholl, I.F.: The precise solar shape and its variability. *Science* **337**, 1638–1640 (2012)

Kurucz, R.L.: New atmospheres for modelling binaries and disks. In: Milone, E.F. (ed.) Light curve modeling of eclipsing binary stars, pp. 93–101. Springer-Verlag, New York (1993)

Leibacher, J.W., Stein, R.F.: *Astrophys. Lett.* **7**, 191–192 (1971)

Leighton, R.B.: Comments on presentation by E. Böhm-Vitense on “Considerations on Localized Velocity Fields in stellar atmospheres: Prototype—the Solar Atmosphere. A. Convection and Granulations: Preview on Granulation—Observational Studies.” *Nuovo Cimento* 22, Series 10, Supplemento, pp. 321–325; discussion, pp. 325–327 (1961)

Leighton, R.B., Noyes, R.W., Simon, G.W.: Velocity fields in the solar atmosphere. *Astrophy. J.* **135**, 474–499 (1962)

Lydon, T.J., Sofia, S.: A measurement of the shape of the solar disk: the solar quadrupole moment, the solar octopole moment, and the advance of perihelion of the planet mercury. *Phys. Rev. Lett.* **76**, 177–179 (1996)

Mann, G., Classen, T., Aurass, H.: Characteristics of coronal shock waves and solar type II radio bursts. *Astron. Astrophys.* **295**, 775–781 (1995)

Maunder, E.W.: Note on the distribution of sun-spots in heliographic latitude, 1874–1902. *Monthly Notices Roy. Astron. Soc.* **64**, 747–761 (1904)

Meyer-Arendt, J.R.: *Introduction to Classical and Modern Optics*, 4th edn. Prentice Hall, Englewood Cliffs, NJ (1995)

Milone, E.F., and Wilson, W.J.F. 2013. *Solar System Astrophysics: Planetary Atmospheres and the Outer Solar System*, 2nd ed. Springer, New York

Nieto, M.M.: *The Titius-Bode Law of Planetary Distances*. Oxford University Press, Oxford (1972)

Pasachoff, J.M., MacRobert, A.: Is the Sunspot Cycle About to Stop? *Sky Telesc.* **122**(3), 12–13 (2011)

Piteva, E.V., Standish, E.M.: Proposals for the masses of the three largest asteroids, the Moon-Earth mass ratio and the Astronomical Unit. *Celestial Mech. Dyn. Astron.* **103**, 365–372 (2009)

Ramesh, R., Shanmugha Sundaram, G.A.: Type I radio bursts and the minimum between sunspot cycles 22 & 23. *Astron. Astrophys.* **364**, 873–875 (2000)

Saint-Hilaire, P., Vilmer, N., Kerdraon. A.: A decade of solar type III radio bursts observed by the Nançay Radioheliograph 1998–2008. *Astrophys. J. Lett.* **747**, L1 (4 pages) (2013)

Schlosser, W., Schmidt-Kaler, Th., Milone, E.F.: *Challenges of Astronomy: Experiments for the Sky and Laboratory* (“*Challenges*”). Springer-Verlag, New York (1991/4)

Thejappa. G., MacDowall, R. J., Bergamo, M., Papadopoulos, K.: Evidence for the oscillating two stream instability and spatial collapse of langmuir waves in a solar type III radio burst. *Astrophys. J. Lett.* **747**, L1 (4 pages) (2012)

Torres, G.: On the use of empirical bolometric corrections for stars. Astrophys. J. **140**, 1158–1162 (2010)

Tousey, R., Milone, E.F., Purcell, J.D., Palm Schneider, W., Tilford, S.G.: *An atlas of the solar ultraviolet spectrum between 2226 and 2992 angstroms*. Naval Research Laboratory, Washington, DC. NRL Report 7788 (1974)

Tun, S.D., Vourlidas, A.: Derivation of the magnetic field in a coronal mass ejection core via multi-frequency radio imaging. *Astrophys. J.* **766** (10 pages) (2013). doi:10.1088/0004-637X/766/2/130

Ulrich, R.K.: The five-minute oscillations on the solar surface. *Astrophys. J.* **162**, 993–1002 (1970)

van Noort, M.J., Ruppe van der Voort, L.H.M.: High-resolution observations of fast events in the solar chromosphere. *Astrophys. J.* **648**, L67–L70 (2006)

Webb, D.F., Howard, T.A.: Coronal mass ejections: observations. *Living Rev. Solar Phys.* **9**, 3–83 (2012)

Weiss, A.A.: The type IV solar radio burst at metre wavelengths. *Aust. J. Phys.* **16**, 526–544 (1963)

Wildt, R.: Negative ions of hydrogen and the opacity of stellar atmospheres. *Astrophys. J.* **90**, 611–620 (1939)

Wolff, C.L.: The five-minute oscillations as nonradial pulsations of the entire Sun. *Astrophys. J. Lett.* **177**, L87–L91 (1972)

Wolff, C.L.: Rotational sequences of global oscillations inside the Sun. *Astrophys. J. Lett.* **580**, L181–L184 (2002)

Wolff, C.L.: Effects of a deep mixed shell on solar g-modes, p-modes, and neutrino flux. *Astrophys. J.* **701**, 686–697 (2009)

Chapter 5
General Properties of the Terrestrial Planets

The terrestrial planets are confined to the inner solar system: Mercury, Venus, Earth (or Earth–Moon), and Mars. They differ from the main belt asteroids (Milone and Wilson 2014, Sect. 15.7.3) in being much more massive, and from the giant planets (Chap. 12) in being much less massive; they also lack the extensive mantles of hydrogen and helium of the gas giants. The rocky, central core of Jupiter is estimated to be between 10 and 20 Earth masses, and, if it became visible, might be considered the principal "terrestrial planet." However, the rocky cores of the outer planets are under greater pressures than are the interiors of the terrestrial planets,[1] and will have different traits. For example, they are expected to be mixed with large quantities of ices, and this is true also of the smaller bodies of the outer solar system. Thus the natures of the planets of the inner solar system are sufficiently unique to be discussed as a separate group, the "inner" or "terrestrial" planets.

5.1 Overview of Terrestrial Planets

The vast bulk of the rocky material of the terrestrial planets is hidden in the interior, so much of what we know about the material of each planet comes mainly from its bulk properties (i.e., the mean density, size, and mass) and the viewable surface. Although only the Moon and Earth have yielded sufficient material for detailed laboratory analyses, the surfaces of Mars and Venus have been examined by on-site probes and the composition has been investigated at a limited number of sites. Mercury is currently (2013) under intense scrutiny by the Messenger orbiter.

[1] Pressures of ~10^5 Pa (1 Pascal = 1 N/m^2 = 10^{-5} bar) for Earth's surface and ~3.64×10^{11} Pa for the Earth's core (Tromp 2001) compare to an estimated 4×10^{10} for the core of Mars (Stewart et al. 2007) and >4×10^{12} for that of Jupiter (Fortney 2007). The physical conditions in the rocky core of Jupiter are beyond current capability to recreate in the laboratory.

E.F. Milone and W.J.F. Wilson, *Solar System Astrophysics: Background Science and the Inner Solar System*, Astronomy and Astrophysics Library,
DOI 10.1007/978-1-4614-8848-4_5, © Springer Science+Business Media New York 2014

On the whole, the major and dwarf planets, and some of the major moons, are more spherical in shape than are most of the minor planets, moons, comets, and meteoroids that permeate the larger solar system, but none of them are precisely spherical. The approximate shape can be determined by direct filar micrometry (from ground-based telescopes) and, in most cases, those have been improved by direct imaging from space craft. The internal mass distributions can be and have been investigated by the perturbations on natural or artificial satellite orbits for all the planets.

The properties of planetary interiors have been investigated primarily through seismic effects (for the Earth and the Moon), magnetic field effects (for some), and the bulk properties such as radii, masses, mean densities, and surface densities (for all). In Chaps. 6 and 7, we deal with the properties of the interior and how they are known; the Moon will be treated separately in Chap. 8.

Telescopic observations of planets have been made since the early seventeenth century, and still continue to provide fresh insights and discoveries. The long history of observations of the once-alleged planet Vulcan, of Mercury and Venus rotation rates from perceived markings, and of putative Martian canals from perceived linear features demonstrates clearly that observations are not foolproof, but the discoveries of the lunar maria, the gibbous phases of Venus, Galilean satellites, the moons of Mars, the planet Uranus, the first asteroid, Ceres, and Charon (the moon of Pluto) changed our perception of the solar system in lasting ways. Ground-based telescopic observations still play a role today. A "great white spot" that suddenly appeared on Saturn and the signature of the first extrasolar planet in high-precision radial velocity measurements in the 1990s were such discoveries.

At the end of this chapter, we will touch on the visibility of planetary surfaces. We will define and discuss the significance of such terms as the *phase,* the *phase angle,* and *phase function* in studying the reflective properties of a solar system object. The surface features of the major terrestrial planets, namely Venus, Earth, and Mars, are quite similar in many respects, but there are also important differences among them, aside from the oceans of water on the Earth and the extensive atmosphere on Venus. In all the planets, impacts have played an important role in determining their physical and dynamical properties. These properties and the similarities and differences among terrestrial planets will be explored more thoroughly in Chap. 9.

Finally, the atmospheres and meteorology of the planets provide interesting differences in chemistry as well as physics. Atmospheric physics and atmospheric and ionospheric chemistry will be treated in detail in Milone and Wilson (2014, Chaps. 10 and 11).

5.2 Bulk Properties

The mean densities of the planets can be obtained simply from the bulk properties of mass and radius. The mass is directly obtained by observation of the semi-major axis of a moon in the case of most of the planets or from the acceleration of a space probe in the case of Mercury, Venus, (most) asteroids, or comets. In either case, the acceleration of the smaller, less massive body is

$$\mathbf{a} = -\mu \hat{\mathbf{r}} / r^2 \tag{5.1}$$

where $\mu = GM$, and M is the planet's mass, r is the distance of the smaller body from the planet's center, and $\hat{\mathbf{r}}$ is a unit vector directed radially away from the planet's center. From (5.1), the magnitude of the acceleration at the surface of a planet of radius R is

$$\mathrm{a}(R) = g(R) = GM/R^2 \tag{5.1a}$$

Given M and R, the planet's mean density follows:

$$<\rho> \quad = \frac{M}{(4/3)\pi R^3} \tag{5.2}$$

The planet with the highest mean density is the Earth (5.515 g/cm^3 or 5,515 kg/m^3), that with the lowest, Saturn (690 kg/m^3). Because different types of material have different densities (e.g., liquid water, ~1,000 kg/m^3; the minerals pyroxene and olivine, ~3,300; iron sulfide, 4,800; and metallic iron, 7,900 kg/m^3), one might suppose that determining the composition would be a simple matter of getting the correct mixture of material. The volume fraction[2] $\mathrm{X_i}$ of material i contributes to the overall density through the relation:

$$<\rho> \quad = \quad \Sigma \rho_i X_i \tag{5.3a}$$

where $<\rho>$ is the mean density.[3] For instance, the planet Mercury has a mean density, $<\rho> = 5{,}430$ kg/m^3. We can predict the fraction of metallic iron, X_{Fe} and some average "rock," of density, say, $\rho = 3{,}500$, with a fraction $X_{\mathrm{rock}} = 1 - X_{\mathrm{Fe}}$. From (5.3a):

$$<\rho> \quad = \quad \rho_{\mathrm{Fe}} X_{\mathrm{Fe}} + \rho_{\mathrm{rock}}(1 - X_{\mathrm{Fe}}) \tag{5.4}$$

from which one may derive[4] $X_{\mathrm{Fe}} = 0.44$, and therefore, $X_{\mathrm{rock}} = 0.56$. But this can be misleading, because the mean density describes the bulk of material that is not at standard temperature and pressure (STP) conditions. Much of it is under extremely

[2] V_i/V_{total} where V is the volume and i a constituent.

[3] One may similarly compute the mass fraction (also referred to as X_i) from a relation of the kind,

$$1/<\rho> = \sum \{X_i/\rho_i\} \tag{5.3b}$$

[4] With the (compressed) mean density, one can derive a mass-fraction, $X_{\mathrm{Fe}} = 0.64$ for metallic iron, assuming that to be the only form that iron takes in Mercury.

Table 5.1 Planetary densities (kg/m^3)

Planet	ρ_{comp}	ρ_{uncomp}	Comments
Mercury	5,430	5,300	Still lots of iron!
Venus	5,240	4,000	
Earth	5,515	4,100	
(*Moon*	*3,360*	*3,300*	*Close to the density of Earth's mantle*)
Mars	3,940	3,700	

high pressure, and associated high temperature, so that the mean density, as determined from the mass and volume, has limited value in predicting the compositional mix. Models have, however, been constructed for planetary structure which permit estimates of the *uncompressed* density. See Table 5.1, taken from Consolmagno and Schaefer (1994, Table 4.2, Ch. 4.1, p. 71).

Notice that the Moon has the lowest mean density among these bodies and that its compressed and uncompressed densities are about the same. Note also the mild trend of lower uncompressed mean densities with increased distance from the Sun. Recalculating the volume and mass fractions with the *uncompressed* mean density of Mercury, the (assumed metallic) iron fractions become 0.41 and 0.61, respectively.

Another observable quantity involving bulk as well as dynamic properties is the *specific angular momentum* (SAM, the angular momentum per unit mass, **h**) of a planet and its satellites, because, as in the solar system generally, the distribution of angular momentum provides important clues to the origin of the planetary and satellite systems.

5.3 Gravitational Potential Fields

In this section we make use of conventional notation, which differs somewhat from that used in earlier chapters. If a planetary mass, M, is spherically symmetric, then a small mass, m, placed at a distance r from the centre of M has a *potential energy* (Fig. 5.1)

$$U = -\frac{GMm}{r} \tag{5.5}$$

The *gravitational potential, V*, is defined by the equivalence

$$\left\{\begin{array}{l}\text{gravitational}\\ \text{potential at } r\\ \text{due to } M\end{array}\right\} \equiv \left\{\begin{array}{l}\text{potential energy per unit}\\ \text{mass for mass } m \text{ located}\\ \text{at distance } r \text{ from } M\end{array}\right\}$$

That is,

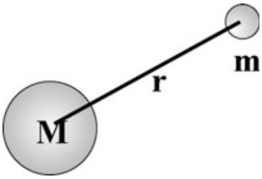


Fig. 5.1 Masses M and m have gravitational potentials and potential energies by virtue of their separation, *r*

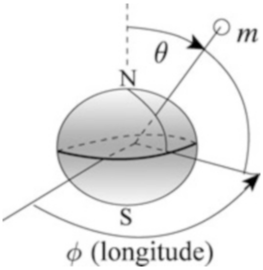


Fig. 5.2 Non-symmetric *M*

$$V \equiv \frac{U}{m} = \frac{\left[-\frac{GMm}{r}\right]}{m} = -\frac{GM}{r}, \tag{5.6}$$

Then, by way of a check, the potential energy of any mass *m* placed in this gravitational potential is:

$$U = mV = m\left(-\frac{GM}{r}\right) = -\frac{GMm}{r},$$

which agrees with (5.5).

If the mass *M* is *not* spherically symmetric, the strength and direction of the gravitational acceleration, **a**, are influenced by the actual distribution of mass inside *M* and by the location of *m* (in terms of r, θ, and ϕ) relative to the center of *M*, as shown in Fig. 5.2. Note the notation convention here, which differs from the usage of spherical astronomy in Chap. 2: the angle θ is called the *co-latitude* and is measured from the pole, not the equator. In the present context, the symbol ϕ is used for the longitude. V is now a function of r, θ and ϕ, and (5.6) no longer applies. It turns out that the equation for V can be written as a summation over an infinite series of spherical harmonics using *Legendre polynomials,* P_n (cos θ), and the *associated Legendre polynomials,* $P_n^m(\cos\theta)$.

These are standard mathematical functions which are often encountered when using spherical coordinates. The properties of these polynomials are summarized in the addendum to this chapter (Sect. 5.6). With them, the potential in this spherical coordinate system can be written:

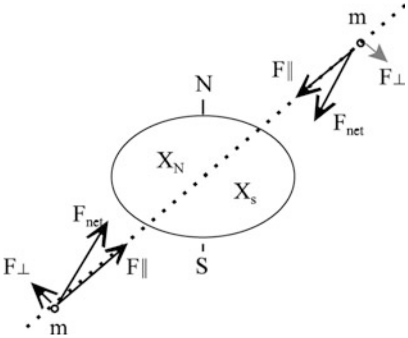


Fig. 5.3 Planet and satellite orbit plane

$$V = -\frac{GM}{r}\left[1 - \sum_{n=2}^{\infty}\left(\frac{a}{r}\right)^{n} J_n P_n(\cos\theta)\right] \\ + \left[\sum_{n=2}^{\infty}\sum_{m=1}^{\infty}\left(\frac{a}{r}\right)^{n}(C_{nm}\cos\ m\phi + S_{nm}\sin\ m\phi)P_n^m(\cos\theta)\right] \tag{5.7}$$

where r is the distance and a is the mean radius of the planet. The polynomials are standard functions, so it follows that the coefficients must be determined by the particular mass distribution of any given planet. If we can find these coefficients from properties of the orbits of moons or of spacecraft, then (5.7) can, in principle, be "inverted" to find the mass distribution inside the planet required to produce these coefficients.

The terms in P_n do not involve the longitude, ϕ, and so describe mass distributions that are symmetric about the rotation axis but vary with latitude (as, for example, in a rotationally flattened sphere). The terms in P_n^m do involve ϕ and so describe the amount of departure from axial symmetry (longitude-dependent mass distributions).

Even-numbered J_n (J_2, J_4,...) describes mass distributions that are symmetric about the equatorial plane (the northern hemisphere is a mirror image of the southern hemisphere), whereas odd-numbered J_n (J_1, J_3,...) describes asymmetric distributions (such as a pear shape, for example). J_1 is missing (i.e., $n = 1$ is missing from the summation) because $n =1$ corresponds to a bodily movement of the sphere in some direction away from the center of the coordinate system. For simplicity, we define our coordinate system to be centered on the center of the sphere, so J_1 disappears. J_2, the coefficient for the second spherical harmonic term, is called the *quadrupole moment* and measures the amount of polar flattening. It is generally the simplest to find. Figure 5.3 shows a satellite (either natural or artificial) in an inclined orbit around a planet which is rotationally flattened.

If there are no other "irregularities" in the planet, then rotation produces a mass distribution which is symmetric about the equator. The plane of the satellite's orbit "slices" the planet into two halves of equal mass, as shown in Fig. 5.3, where the dotted line marks the plane of the orbit of the small body of mass m and where X_N and X_S mark the centers of mass of the northern and southern portions of the planet delineated by the plane of the orbit. When the satellite is north of the planet's equator, it is closer to the center of mass of the southern half of the planet X_S than to that of the northern half, X_N, as shown, and so feels a net off-center pull toward the south. This force can be resolved into two components, one in the plane of the orbit (toward the planet's center) and one at right angles to the orbit. The force toward the center is the centripetal force, the geometric description of the gravitational force of the planet that keeps the satellite in orbit, while the force at right angles in effect creates a torque on the orbit. As with a simple gyroscope spinning on one end on a desk, this gravitational torque causes the orbit to precess. The rate of precession depends on the amount by which the center of mass of each half of the planet is offset from the planet's center, and therefore the precession rate also depends on the amount of flattening and on the mass distribution inside the planet (the same amount of flattening will produce less offset in a planet with mass more concentrated toward the center).

J_2 can be calculated from the rate of precession and provides an important tool for probing the interior of a planet.

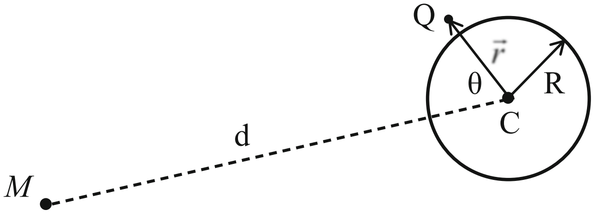
The higher-order terms and longitude-dependent terms are due to departures from the simple rotational flattening described above and give a measure of the extent to which non-hydrostatic forces (i.e., other than gravity and rotation; for example, rising mantle plumes, mountain-building due to lithospheric convergence, massive lava flows supported by stresses in the planet, etc.) have affected the mass distribution in the planet. They show up observationally as changes in the plane, orientation, size, eccentricity, and period of the satellite's orbit.

Two other useful sets of numbers, h_n and k_n ($n \geq 2$), were developed by A. E. H. Love (1911) to describe tidal effects (see Sect. 3.7). Specifically, h_2 describes the ratio of the height of the solid-body tide to the height of the static equilibrium tide (an idealized oceanic tide in which the water achieves static equilibrium with the tidal forces, forming a prolate spheroid with its long axis in line with the perturbing body), and the *potential Love number* k_2 describes the ratio of the gravitational potential of the perturbed (tidally-deformed) body at a point to the perturbing potential at that point. Here, we are concerned with k_2.

Consider an initially-spherical satellite of radius R whose distance, d, from a planet of mass M is small enough that the satellite is tidally distorted by the planet, but large enough that the planet can be treated as a point mass, as illustrated in Fig. 5.4. The planet-satellite scenario is used here for convenience of terminology, but the arguments apply equally well to any two bodies that satisfy the conditions in the previous sentence; e.g., a planet tidally-distorted by its satellite, a planet orbiting the Sun or another star, or a planet being perturbed by another planet.

The gravitational potential, $\Phi(\vec{r})$, at any point Q located at a displacement $\vec{r}$ from the center, C, of the satellite in Fig. 5.4 can be written as

Fig. 5.4 A figure for defining the potential Love numbers



$$\Phi(\vec{r}) = W(\vec{r}) + V(\vec{r}) \tag{5.8}$$

where $W(\vec{r})$ is the potential at Q due to the planet and $V(\vec{r})$ is the potential at Q due to the satellite. If a spacecraft or other object passes the satellite, then $W(\vec{r})$ determines the path that the spacecraft would follow if the satellite were not there, and $V(\vec{r})$ determines how this path is perturbed by the presence of the satellite.

$W(\vec{r})$ can be expanded in a harmonic series using the Legendre polynomials defined above:

$$W(\vec{r}) = -\frac{GM}{d}\sum_{n=0}^{\infty}\left(\frac{r}{d}\right)^n P_n(\cos\theta) \equiv \sum_{n=0}^{\infty} W_n \tag{5.9}$$

It is instructive to examine the first three terms of (5.9):

$$\begin{aligned} W(\vec{r}) &= -\frac{GM}{d}\left(\frac{r}{d}\right)^0 P_0(\cos\theta) - \frac{GM}{d}\left(\frac{r}{d}\right)^1 P_1(\cos\theta) - \frac{GM}{d}\left(\frac{r}{d}\right)^2 P_2(\cos\theta) - \dots \\ &= -\frac{GM}{d}(1)(1) - \frac{GM}{d^2}r\cos\theta - \frac{GM}{d^3}r^2\frac{1}{2}(3\cos^2\theta - 1) - \dots \end{aligned} \tag{5.10}$$

The n = 0 term is the gravitational potential due to the planet at the center of the satellite, the n = 1 term is related to the gravitational acceleration due to the planet at the center of the satellite, and, from (3.55), the n = 2 term is related to the tidal force per unit mass on the satellite by the planet. The *tide-raising potential* is thus given by the sum of terms n $\geq$ 2 in (5.9). Because of the factor $(r/d)^n$ in (5.9), if r << d then terms n > 2 (i.e., terms beyond those shown in (5.10)) can often be neglected.

We note here that the n = 1 term in (5.9) and (5.10) is present because the origin of the coordinate system is at the center of the satellite and not at the center of the gravitating body (the planet); compare this with (5.7), where the n = 1 term is absent (the summation begins at n = 2) because the coordinate system is centered on the gravitating body.

$V(\vec{r})$ can in principle be found by adding up (i.e., integrating) the potentials at Q due to all the mass elements, dM, within the satellite:

$$V(\vec{r}) = -G\int \frac{dM}{|\vec{r} - \vec{r}'|} = -G\int \frac{\rho(\vec{r}')}{|\vec{r} - \vec{r}'|}\mathrm{dv}' \tag{5.11}$$

where $\rho(\vec{r}')$ is the density at any point $\vec{r}'$ within the satellite, and dv′ is the volume element at point $\vec{r}'$. $V(\vec{r})$ thus depends on the density distribution within the satellite. In practice, the density distribution within the satellite is unknown, and we would like to learn something about it by evaluating the tidal terms in (5.9).

The potential at Q due to the satellite can be described by a harmonic series of the same form as (5.9). Love (1911, p. 53) showed that the tidally-induced potential V_2 at r can be represented by a function K(r) times the tide-raising potential, W_2, from (5.9). Then, neglecting terms $n > 2$,

$$V = V_0 + V_2 = V_0 + K(r)W_2 \tag{5.12}$$

where V_0 is the potential at r of the undistorted (spherical) satellite, the (missing) second term from (5.9) is zero because the coordinate system is centered on the gravitating body for potential V (the satellite), and the function K(r) is constant over a surface of a given r.

The potential Love number k_2 is defined from (5.12) as the ratio of the tidally-induced potential to the tidal potential, both evaluated at the mean radius, R, of the satellite; i.e., $k_2 \equiv K(r = R)$:

$$k_2 \equiv \frac{V_2(r = R)}{W_2(r = R)} \tag{5.13}$$

(If terms $n > 2$ are retained then there are also higher-order potential Love numbers, k_n.)

For a uniform satellite, k_2 is related to the *effective rigidity*,

$$\widetilde{\mu} = \frac{19\mu}{2\rho g R} \tag{5.14}$$

of the satellite by

$$k_2 = \frac{3}{2}\left(\frac{1}{1 + \widetilde{\mu}}\right) \tag{5.15}$$

(Love 1911, p. 62; see also Henning et al. 2009). Here, μ is the rigidity of the material, ρ is the density, and g is the gravitational acceleration at the surface of the satellite (r = R). The quantity μ is discussed again in Sects. 5.4.1 and 5.4.2. The effective rigidity, $\widetilde{\mu}$, is a dimensionless ratio that compares the elastic forces in the numerator to the gravitational forces in the denominator (Henning et al. 2009); i.e., $\widetilde{\mu}$ is a measure of how effective the rigidity is at preventing the body from reaching hydrostatic equilibrium by self-gravitation. The product $\rho g R$ can be thought of as a "gravitational stiffness" analogous to the rigidity. If $\widetilde{\mu} >> 1$ the satellite is dominated by its rigidity and if $\widetilde{\mu} << 1$ it behaves like a self-gravitating fluid.

From (5.15), k_2 can range from 0 for $\widetilde{\mu} = \infty$ to 1.5 for $\widetilde{\mu} = 0$. The former applies to a completely rigid satellite [this can also be seen from (5.13): if the satellite remains tidally undistorted, then $V_2 = 0$] and the latter to an incompressible fluid body. A satellite with a fluid layer, as, for example, a subsurface ocean, would typically have a value of k_2 near 0.4. Thus, although (5.11) cannot be inverted to find the density distribution once k_2 is known, the value of k_2 does provide useful constraints on interior models of satellites orbiting planets, or planets orbiting stars.

In practice, k_2 is generally determined from the motion of a spacecraft as it passes by or is in orbit around the satellite.

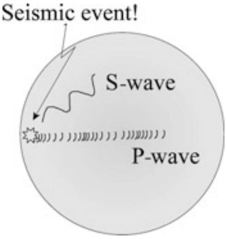
5.4 Structure of the Earth

Through the physics of wave phenomena, the structure of the interiors can be explored. The type and amount of refraction of waves are the means of exploration. Seismic events can be either natural ("quakes") or artificial, i.e., an explosive charge can be set off at some site and the response as a function of distance from that site can be recorded by seismometers set up in a net around the site. In particular, changes in wave velocity show up in a distribution of travel times with distance. The delay time as a function of position along the surface can be modeled by gradual and sharp changes of velocity in the interior and can reveal the presence of discontinuities in the refractive properties of the interior layers.

5.4.1 Seismic Studies

Seismic studies are most mature for the Earth, but the Moon has an array of seismic detectors on its surface which were placed there by the Apollo astronauts between 1969 and 1972. Thus far there have been no comparable studies of Venus and Mars. On Venus, Veneras 13 and 14 carried seismometers that returned data during the 1–2 h that each lander operated on the surface. No events were detected by Venera 13, and Venera 14 detected two possible microseismic events from sources estimated to be within 3,000 km of the lander (Ksanfomaliti et al. 1982); however, while considered less likely, an origin in wind gusts or other local phenomena could not be entirely ruled out. The only seismometers placed on Mars were on the Viking 1 and 2 landers, but the one on Viking 1 did not deploy properly and was unable to return

Fig. 5.5 P- and S-waves



data. The Viking 2 seismometer operated for 500 Martian days, during which one candidate local seismic event was identified (Lorenz and Nakamura 2013). No future lander missions have been announced for Venus, although Hunter et al. (2012) have developed a high-temperature seismometer for use on Venus should such a mission occur. NASA's InSight lander mission to Mars, currently scheduled for launch in 2016, will carry a seismometer and other geophysical instrumentation. It was originally hoped that ESA's ExoMars mission would include a seismometer, but this and several other instruments were cancelled in the final design selection.

Seismic events such as earthquakes release energy that displaces and distorts rock at considerable distances primarily through two types of waves: *P-waves* and *S-waves,* illustrated in Fig. 5.5.

P or *primary* waves are also known as pressure waves, push-pull waves, longitudinal waves, and compressional waves. They are analogous to sound waves.

S or *secondary* waves are also known as shear waves, shake waves, and transverse waves. These are analogous to light or, more generally, electromagnetic waves.

P- and S-waves are *body waves* (through the Earth). *Surface waves* are also produced: L-waves and Rayleigh-waves, analogous to water waves and ripples, respectively. These diminish rapidly below the surface.

When waves are generated by an earthquake, the P-waves travel faster than the S-waves and arrive first at any given location. The speed of a wave depends on:

1. The type of wave (P or S)
2. The density of the medium
3. The state of the material (solid, liquid, or gas)
4. The composition (granite, basalt, iron, . . .)
5. The mineral phase (e.g., the wave speed is different in graphite and diamond, even though they are both forms of carbon)
6. The compressibility, K, for P-waves, and the rigidity, μ, for P- and S-waves. K and μ depend on the density, and the density, in turn, depends on the composition and the mineral phase (as well as the temperature and the pressure), so in fact the wave speed depends very strongly on the density

Because pressure ("P") and shear ("S") waves differ in their speeds through material (e.g., the S-waves do not propagate through fluids), the physical state of the material can be found. Thus the Earth has an outer core that is liquid, revealed by a "shadow zone" in which the S-waves are not seen on the surface. The structure of the interior is then deduced through the solution of equations describing the

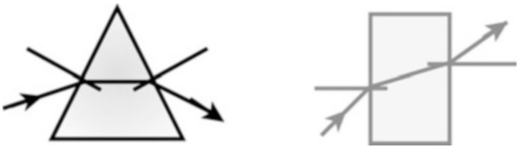


Fig. 5.6 Snell's law

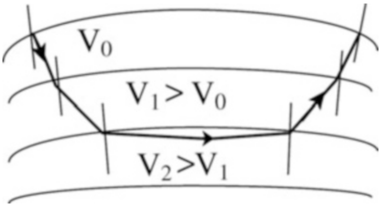


Fig. 5.7 Constant *v* in each layer

expected physics of the interior: an equation of state, which relates the density, pressure, mean molecular weight, and temperature; an equilibrium equation that relates the pressure with the weight per unit area; expressions of the conservation of mass; and an expression which describes heat flow.

The travel paths of seismic waves are decided by the refraction properties of the medium. Snell's law applies:

$$n_1 \sin\theta_1 = n_2 \sin\theta_2 \tag{5.16}$$

where the index of refraction, *n*, depends on the speed of the wave.

As Fig. 5.6 demonstrates, a wave is bent *toward* the normal if it travels from a region of lower refractive index to a region of higher refractive index (i.e., higher wave speed to lower wave speed), because $n_2 > n_1 \Rightarrow \theta_2 < \theta_1$. It is bent *away from* the normal if it travels from a region of higher refractive index to a region of lower refractive index (corresponding to a region with lower wave speed to one of higher wave speed), because

$$n_1 > n_2 \quad \Rightarrow \quad \theta_2 > \theta_1.$$

The path in a layered sphere involves additional considerations. Density generally increases with depth into the Earth so, except at certain discontinuities, the wave speed, *v*, increases with depth. The refractive index thus generally decreases with depth. We consider five conditions:

1. A layered sphere, *v* constant in each layer (Fig. 5.7): the ray refracts *away* from the normal going in and *toward* the normal going out.
2. A sphere in which *v* increases smoothly with depth (Fig. 5.8): this is the same as if the layers were infinitesimally thin. The refraction then becomes continuous.
3. A seismic discontinuity at which *v* suddenly decreases and afterward continues to increase (Fig. 5.9): the refractive index suddenly increases, so the wave refracts toward the normal.

   Also, both reflection and refraction take place at any boundary.

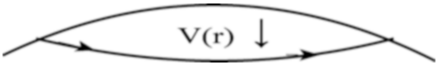


Fig. 5.8 v increasing smoothly with depth

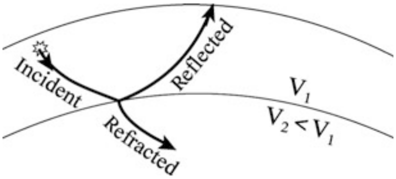


Fig. 5.9 v increase discontinuous

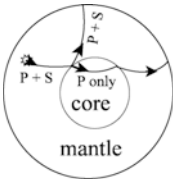


Fig. 5.10 Liquid layer passes only P-waves

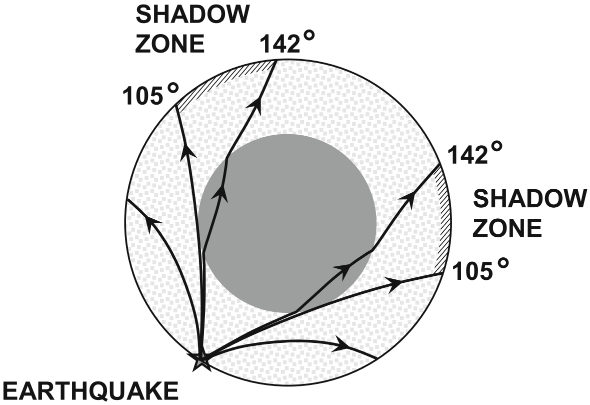


Fig. 5.11 $v_{\text{liq core}} << v_{\text{mantle}}$. Waves that just miss the core boundary return to the surface 105° from the earthquake, whereas those that just contact the core boundary refract into the core and out again, and return to the surface 142° from the earthquake. Between the two is a shadow zone in which no direct P waves are seen

4. Liquid layer: S-waves cannot propagate unless there is a restoring force in the transverse direction. No such force exists in a liquid (a liquid has no "rigidity"), so S-waves cannot propagate in a liquid. No S-waves are observed to travel through the outer core, so the outer core is known to be liquid.

   It appears to be molten iron (Fig. 5.10).
5. Shadow zone (see Figs. 5.11 and 5.12): The speed of P-waves in the liquid outer core of the Earth is much less than in the mantle just above it (13 km/s in the mantle, 8 km/s in the core); v thus decreases suddenly at the core–mantle boundary.

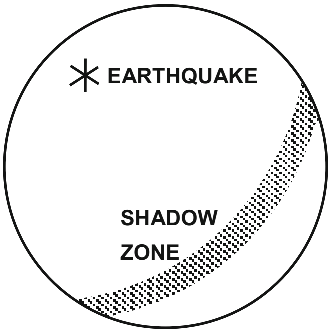


Fig. 5.12 No seismic waves in shadow zone

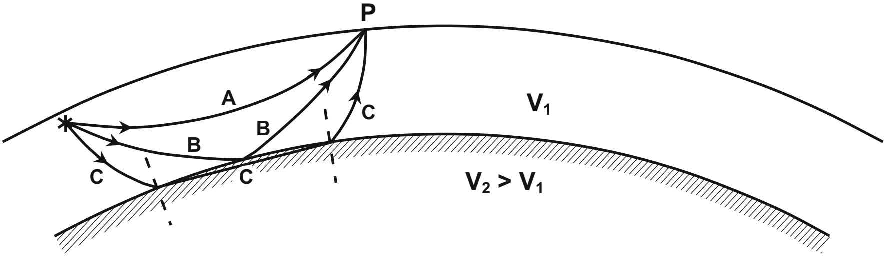


Fig. 5.13 Three wave paths from earthquake site to Station P

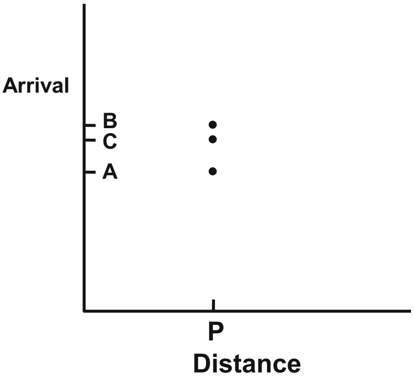
As shown in Figs. 5.11 and 5.12, refraction at the core-mantle boundary creates a shadow zone where no seismic waves will be received. (Body waves may reach there by other processes, such as reflection from the inner core, but these are much weaker than the "direct" P-waves so the shadow zone is still easily observable.) The speed of the wave determines its behavior, so the time intervals for the wave to reach various locations from the site of origin provide the data for modeling the refractive properties of the layers through which it passes.

If there is a seismic discontinuity in which *v* increases suddenly at some depth, then a station located within a certain range of distances from an earthquake can receive waves by three different travel paths (Fig. 5.13):

Speed increases smoothly with depth until the discontinuity is reached, so the station receives waves as follows.

A. Seismic wave takes the most direct path. This has the slowest average *v* but the shortest path; it arrives first.
B. Seismic wave reflects from the discontinuity. It has faster average speed (travels deeper) but along a longer path; it arrives later than A.
C. The wave refracts into the faster layer, then out again. It has a longer path than B, but the higher speed in the lower layer more than compensates for the longer path; it arrives before B, but after A. The arrival times at station P are plotted in Fig. 5.14.

Fig. 5.14 Arrival times of seismic waves at station P



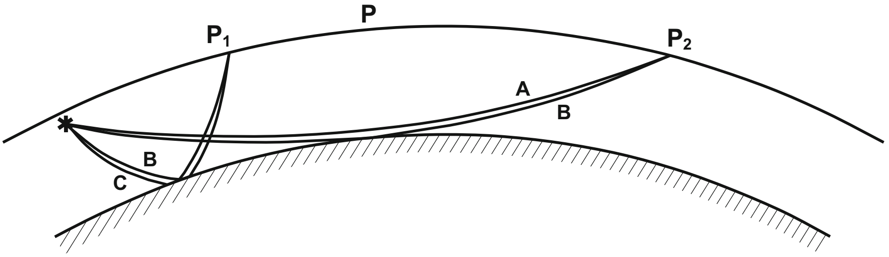


Fig. 5.15 Moving station P change the travel times

If we move P closer to the earthquake site, then the travel path of wave C in the lower layer decreases, eventually becoming zero at point P_1 in Fig. 5.15. The travel times for waves B and C thus approach each other, becoming equal at P_1 .

If we move P further from the earthquake, then the lengths and travel times for paths A and B approach each other, becoming equal at point P_2 in Fig. 5.15.

If we plot the arrival times of waves observed at various stations as a function of the distance of each station from the epicenter of the earthquake (the point on the surface of the Earth directly above the earthquake), then we find the arrival time plot as shown in Fig. 5.16

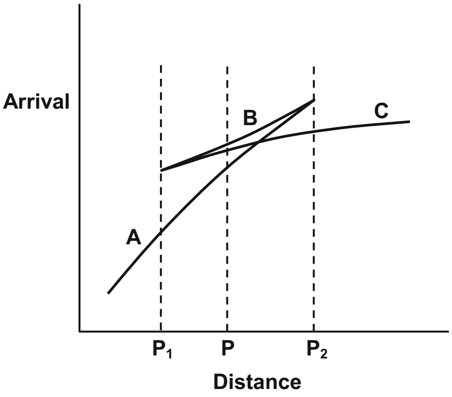
Plots such as these can be used to find wave speed as a function of depth into the Earth.

An example of seismic waves arriving at stations in Alberta, Canada, from the Fukushima, Japan, earthquake of March 11, 2011, is shown in Fig. 5.17. This earthquake measured magnitude 8.9 on the Richter scale (Richter 1935; Gutenberg and Richter 1936).

5.4.2 *The Adams–Williamson Equation*

A derivative expression, the Adams–Williamson equation, relates the velocities of the P and S-waves deduced from the seismic models to the density gradient, $d\rho/dr$,

Fig. 5.16 Arrival time of S-waves as a function of the distance of the observing station from the earthquake, measured along the surface of the Earth. The data are from a simplified two-layer model of the outer mantle with a seismic discontinuity 670 km below the surface, and S-wave speeds 4.0 km/s above the discontinuity and 6.0 km/s below



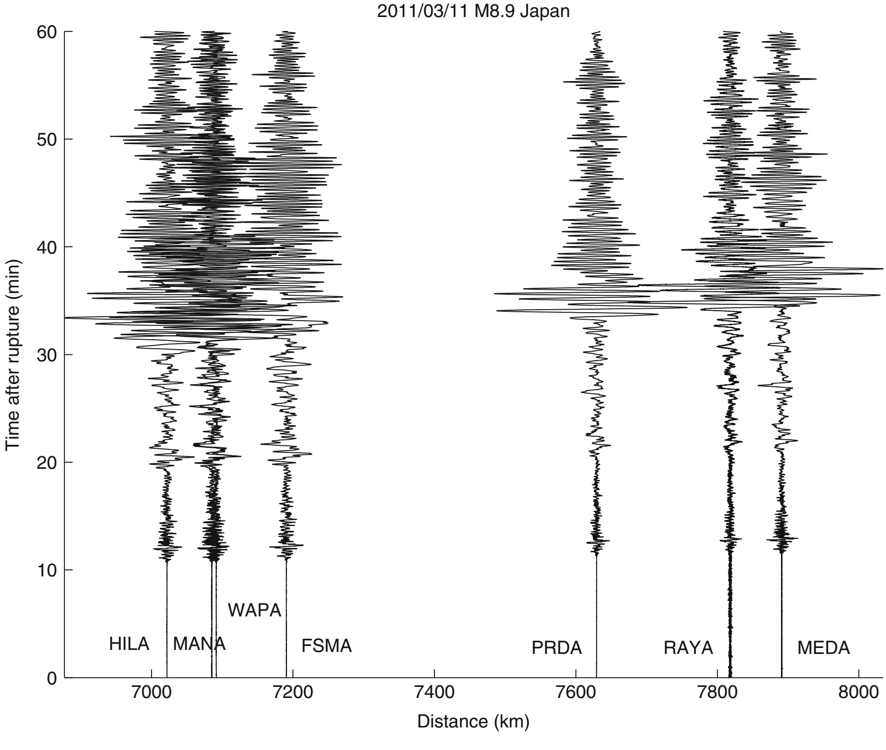


Fig. 5.17 Seismograms of the magnitude 8.9 earthquake in Fukushima, Japan, on March 11, 2011, recorded at seven seismograph stations in Alberta, Canada. The x-axis shows distance from the epicenter. Time increases upward. The initial, lower-amplitude, part of each waveform represents body (P and S) waves that have travelled on shorter paths through the mantle. Of these, the P waves arrive first, followed by the S waves. The later, high-amplitude waves that arrive 30 s or more after rupture are surface waves that have travelled close to the Earth's surface from the earthquake to the seismometer station. FSMA = Fort Smith, HILA = High Level, MANA = Manning, MEDA = Medicine Hat, PRDA = Rothney Astrophysical Observatory near Priddis, RAYA = Raymond, WAPA = Wapiti River. Courtesy, Dr. David Eaton, University of Calgary Dept. of Geoscience

at a distance r from the center. The usual procedure is to consider the interior to be a series of shells of thickness dr and to express the change in density, dρ, across that shell. The density depends also on the molecular weight of the material in the interior, and this is not known exactly. However, it can be guessed. The Adams-Williamson equation is derived as follows:

Start with the equation of hydrostatic equilibrium (the derivation of which is one of your end-of-chapter Challenges!):

$$\frac{\mathrm{d}P}{\mathrm{d}r} = -\rho g = -\frac{GM(r)}{r^2}\rho \tag{5.17}$$

where $M\ (r)$ is the mass enclosed within radius r.

Define the *bulk modulus*, k, through the relation

$$\frac{\mathrm{d}\rho}{\mathrm{d}P} = \frac{\rho}{k} \tag{5.18}$$

Rewriting (5.18), substituting from (5.17), and dividing top and bottom of the right-hand side by ρ, we get:

$$\frac{\mathrm{d}\rho}{\mathrm{d}r} = \frac{\mathrm{d}\rho}{\mathrm{d}P}\frac{\mathrm{d}P}{\mathrm{d}r} = \frac{\rho}{k}\frac{\mathrm{d}P}{\mathrm{d}r} = -\frac{GM(r)\rho}{\frac{k}{\rho}r^2} \tag{5.19}$$

Then, making use of the relations between these quantities and the primary and secondary wave speeds,

$$v_p^2 = \frac{m}{\rho} = \frac{k + \frac{4}{3}\mu}{\rho}, \quad v_s^2 = \frac{\mu}{\rho} \tag{5.20}$$

where m is the *elastic modulus* and μ is the *rigidity* or *shear modulus*,[5] so that, combining the (5.20) expressions,

$$v_p^2 = \frac{k}{\rho} + \frac{4}{3}v_s^2 \tag{5.21}$$

and, rearranging (5.21),

$$\frac{k}{\rho} = v_p^2 - \frac{4}{3}v_s^2 \tag{5.22}$$

[5] For the derivation of these quantities in terms of stresses and Young's modulus, see, for example, Stacey (1969), pp. 85–86.

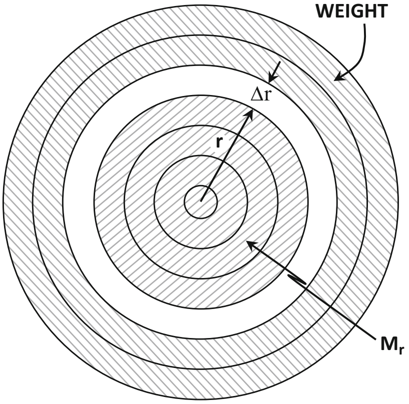


Fig. 5.18 Shells for numerical integration of (5.25)

we arrive at what has been called the "fundamental equation of geology," the *Adams–Williamson* equation:

$$\frac{\mathrm{d}\rho}{\mathrm{d}r} = -\frac{GM(r)\rho}{r^2\left(v_p^2 - \frac{4}{3}v_s^2\right)} \tag{5.23}$$

Now we go on to discuss how to solve this equation in order to find the density, ρ, as a function of *r* inside the Earth.

Equation (5.23) tells us the change ($\mathrm{d}\rho$) in ρ over some small distance d*r*:

$$\frac{\mathrm{d}\rho}{\mathrm{d}r} = f \qquad \Rightarrow \qquad \mathrm{d}\rho = f\,\mathrm{d}r \tag{5.24}$$

where *f* represents the right-hand side of (5.23). Then to find ρ as a function of *r*, we need to integrate (5.24).

$$\rho = \int \mathrm{d}\rho = \int f\,\mathrm{d}r. \tag{5.25}$$

This involves numerical integration; we describe the procedure in a step-wise manner:

First, divide the Earth into many thin shells, as shown in Figs. 5.18 and 5.19.

Each shell has uniform composition and density.

Then the density, ρ, in each shell depends on:

1. mineral composition;
2. compression of the shell by the weight of all the shells above it.

Decide on a model for the Earth's composition as a function of *r*, for example: olivine/pyroxene mantle, iron core.

Note that we can express the total mass, M_r, interior to any radius, *r*, either by:

$$M_r = \int_{a=0}^{r} \mathrm{d}m = \int_{a=0}^{r} \rho(a)\,\mathrm{d}V = \int_{a=0}^{r} \rho(a)\,4\pi a^2\ \mathrm{d}a \tag{5.26}$$

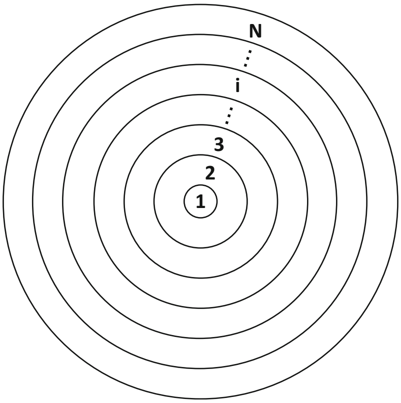


Fig. 5.19 Setting up *N* Shells

or by

$$M_r = M_E - \int_{a\,=\,r}^{R_E} \,\mathrm{d}m = M_E - \int_{a\,=\,r}^{R_E} \rho(a)\,4\pi a^2 \;\,\mathrm{da} \tag{5.27}$$

where M$_E$ and R$_E$ are the Earth's mass and radius respectively. Equation (5.27) is more useful than (5.26), because it involves only quantities above the radius *r*, and these are known at any given *r*, as described below.

As we are going to integrate numerically, we need numbers, and an infinitesimal quantity like d*a* or d*r* is not a number. Therefore, we need to convert (5.23) and (5.27) to *finite difference equations*, the mathematical equivalent of "going digital."

Taking (5.27) first, we replace d*a* by the width, Δr_i, of the i$^{\mathrm{th}}$ shell:

$$M_r = M_E - \sum \rho_i 4\pi r_i^2 \;\, \Delta r_i \tag{5.28}$$

We know M_{E} and R_{E}, and we know ρ at the surface of the Earth. Now we can go through an iterative loop, working inwards from the Earth's surface.

LOOP:

1. Decide on a thickness Δr_i for this shell.
2. Use (5.28) to calculate M_r (Note: the RHS is known). This involves numerical integration.
3. Rewrite the Adams–Williamson equation as a finite difference equation:

$$\Delta\rho_i = \left\{ \frac{GM_r\rho_i}{r_i^2\left[(v_P)_i^2 - \frac{4}{3}(v_S)_i^2\right]} \right\} \;\Delta r_i \tag{5.29}$$

where M_r was found above, and v_P and v_S are known from the seismic data.

This gives $\Delta\rho_i$, the change in ρ across the shell, so ρ for the next shell inward is:

$$\rho_{i-1} = \rho_i + \Delta\rho_i. \tag{5.30}$$

We continue looping until we either run out of mass ($M_r = 0$) or run out of radius ($r_i = 0$). Because M_r has to be 0 exactly at $r = 0$, we have to adjust the composition and other input assumptions until they both go to zero at the same time. We now have the density, ρ, as a function of r. The model also has to allow for mineral phase changes, e.g., the crystal structure of olivine changes at about 150–200 kbar pressure to a new crystal structure; it is then called *spinel* and is about 10 % denser than the original crystal structure (*olivine*) at the same pressure. See Chap. 7 for a discussion of mineral structure.

5.4.3 Moments of Inertia

The moment of inertia I, of the Earth depends on the distribution of mass (and therefore density) inside the Earth:

$$I = \sum m_i r_i^2 \tag{5.31}$$

For a uniform sphere of radius R and mass M, one may show that

$$I = (2/5)\ MR^2 \tag{5.32}$$

In general, however,

$$I = KMR^2 \tag{5.33}$$

See Fig. 5.20 for the effect of different density distributions on K. Note that I is lower if most of the mass is close to the center (in the core) and I is higher if most of the mass is further out (in the mantle). We can calculate the moment of inertia for our model of density as a function of radius and compare that to the observed value. If the two are different then we can adjust our model to bring it into agreement with the observations.

With regard to orbiting satellite observations, we note that the second harmonic can be written in terms of the moments of inertia:

$$J_2 = \left[I_z - I_{x,y}\right] / \left(M r^2\right) \tag{5.34}$$

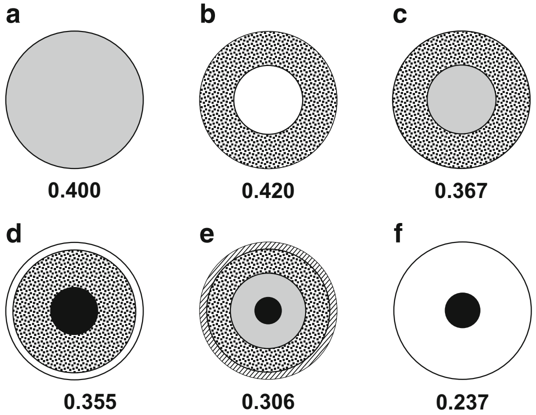


Fig. 5.20 The effect of density distribution on K. Outer radius, 1 unit. (**a**) *Undifferentiated asteroid or comet.* Uniform density. (**b**) *Density inversion.* Core: $r = 0.5$, $\rho = 0.5$; mantle: $\rho = 1.00$. (**c**) Core: $r = 0.5$, $\rho = 2.00$; mantle: $\rho = 1.00$. (**d**) *Callisto model.* Core: $r = 0.34$, $\rho = 3.5$; mantle: $r = 0.89$, $\rho = 2.1$; crust: = 1.0. (**e**) *Terrestrial planet model.* Inner core: $r = 0.19$, $\rho = 12.8$; outer core: $r = 0.55$, $\rho = 11.2$; lower mantle: r = 0.895, $\rho = 4.9$; upper mantle: $\rho = 3.6$. (**f**) *Jovian planet model.* Core: $r = 0.25$, $\rho = 20$; mantle: $\rho = 0.4$. (Densities in g/cm^3.)

where I_z is the moment of inertia about the rotation axis of the planet and $I_{x,y}$ is the moment of inertia about an axis in the equatorial plane. Two effects of J_2 on the orbit (see Sect. 3.2 of Chap. 3 for definitions) are:

$$\mathrm{d}\omega/\mathrm{d}t = -\ (3/2)J_2\left(n/p^2\right)\left[(5/2)\ \sin^3 i - 2\right] \tag{5.35}$$

the change in the argument of pericenter, where $p = \mathrm{a}(1 - \mathrm{e}^2)$, and

$$\mathrm{d}\Omega/\mathrm{d}t = -\ (3/2)J_2\left(n/p^2\right)\ \cos i \tag{5.36}$$

the change in the longitude of the ascending node. An effect on the mean motion, and thus the position of the object in its orbit, is:

$$[t - T_0]\,\mathrm{d}n/\mathrm{d}t = -(3/2)J_2\left(n/p^2\right)\left[(5/2)\ \sin^3 i - 2\right] \tag{5.37}$$

5.4.4 Models of the Interior

Knowledge of the Earth's interior can be obtained, with varying degrees of certainty, from seismology, geomagnetism, geochemistry, high-pressure experiments, and samples of crustal material. To some extent, the mantle can even be sampled directly: some upper mantle material reaches the surface in mid-ocean ridge lavas, and some lower-mantle material is believed to become entrained in deep mantle plumes and reaches the surface in some ocean island basalts (hot-spot volcanism) (Deschamps et al 2011). Cosmically, a rare class of carbonaceous meteorite, the CI chondrites [five examples known from falls; a few more are known from finds in Antarctica (Islam et al. 2012)], have almost exactly solar abundances of the elements, with the exception of hydrogen and helium. They are therefore believed to be remnants of the original protoplanetary disk (i.e., they are the most primitive meteorites), and provide a reference for studying processes in the Earth that have altered these abundances. For example, assuming a bulk CI chondritic composition for the Earth, *lithophile* ("rock-loving") elements that are over-abundant in the mantle would be so because they were excluded from the core, and *siderophile* ("iron-loving") elements that are depleted in the mantle would be so because they were preferentially carried into the core. Iron, stony-iron and some stony meteorites also provide information about *differentiation*, the separation of planetary and asteroidal bodies into crust, mantle and core. Knowledge of the present state of the Earth's interior then provides a basis for modeling the formation and evolution of the Earth. For a comprehensive discussion of these meteorites, see Chap. 15 in Milone and Wilson (2014), for chondrites in particular, Sects. 15.1.3.1 and 15.2, and for CI especially, 15.2.2.

On the computational side, numerical integration of the Adams–Williamson equation produces a march of the density, ρ, with distance from the center (Sect. 5.4.2). The models are constrained by the surface boundary layer where the density, pressure, and temperature drop sharply, and by the bulk properties of the planet: the mean density and radius. Once $\rho(r)$ is known, equations of state (equations relating T and P to ρ) can be used to find T and P as functions of r. Examples of models may be seen in Figs. 5.24, 5.25, and 5.26 and 6.12. In these models, density, temperature and pressure are all seen to increase inward through the interior.

Radioactive dating of meteorite samples has provided an age for the solar system of 4.568×10^9 year (Wood 2011). This is the time at which dust in the solar nebula began to accrete to form the planets, asteroids and comets. The first small objects formed by dust grains clumping together, and as objects became more massive, collisions resulting from mutual gravitational perturbations became more important. By ~10^4–10^5 years, when many objects were in the 10-km size range, gravitational perturbations dominated the accretion process, and by ~10^6 years a few tens of objects had grown into planetary embryos of a few per cent the mass of the Earth.

The growth of the terrestrial planets appears to have followed an approximately exponential law, $M(t) = M_{final}(1 - e^{-t/\tau})$, with a time constant, τ, in the case of the Earth of approximately 10 or 11 million years (Wood 2011). Thus, growth was most rapid at the start, 95 % complete at $t = 3\tau$, and 98 % complete at 4τ (~40 Ma for the Earth).

The Earth's core (discussed in detail below) is expected to contain primarily iron with substantial amounts of nickel, mixed with sulfur, oxygen, and other elements. The mantle and crust of the Earth, on the other hand, are much richer in silicates. With an assumption about the density at the center ($r = 0$), the differences in density can be added, shell by shell, to determine the interior density at every radius from the center. In this way, the interior structure can be induced. Thus the core takes up the inner 3,500 km from the center (the outer 2/3 of which appears to be liquid), the mantle approximately 2,900 km, and the crust a mere 50–100 km. The mantle itself has structure: in the upper mantle, a deeper, more fluid component (the *asthenosphere*) underlies the solid *lithosphere,* which includes the crust.

The lithosphere "rides" on the asthenosphere in the form of "plates" driven by convective regions in the asthenosphere. The effect is to produce regions of *orogenesis,* or mountain-building, and *subduction,* regions where one plate (e.g., an oceanic plate) flows below another (typically a continental plate). The result is continental drift, in which continents move at rates of the order of 1 cm/year, and where areas which are subducted are replaced by new influx of material along mid-oceanic rises, producing seafloor spreading. The vent areas are outgassing, warm, and rich in sulphur and other material, and it is suspected that life began on the Earth in just such areas; in any case, there is a variety of living organisms currently found at such sites. The temperature of the lava that spreads out away from the upwelling areas rapidly falls below the *Curie point*, so the ferro-magnetic material in it aligns with the local magnetic field in effect at the time and remains that way. Thus, the magnetic reversals in the Earth's field can be studied in the remanant magnetism trapped in successive bands of cooled lava oriented parallel to the mid-ocean ridge as the ocean floor spreads away from the ridge.

The major arguments supporting iron as the dominant component of the Earth's core are,

1. Density models derived from seismic data and the Adams-Williamson equation (and therefore independent of assumptions about composition) predict densities from 10,000 kg/m^3 at the core boundary to 13,000 kg/m^3 at the centre; see, e.g., Fig. 5.25. Silicates cannot be compressed to this density by the pressures expected in the Earth's core.
2. The fluidity and electrical conductivity of the silicates in the Earth's mantle are too low to produce the observed magnetic field of the Earth. The magnetic field is therefore produced by the core. This requires the core to be both liquid and metallic.
3. The seismologically observed density and P-wave velocity (sound speed) in the core are close to those of iron measured in the laboratory at similar temperatures and pressures. (These temperatures and pressures can be produced in the lab in a special device called a diamond-anvil cell.)
4. Cosmically, iron is by far the most abundant element having these properties.

An important insight into core formation in the Earth is provided by the measured Ni/Co (nickel to cobalt) ratio in the mantle. Ni and Co are moderately siderophile and both are depleted in the mantle relative to chondritic composition, as expected. However, at least at low pressures, Ni is more strongly siderophile than Co and should be more depleted, whereas in fact the Ni/Co ratio in the mantle is close to chondritic. The solution to this discrepancy was provided by high-pressure experiments (Li and Agee 1996) which show that both Ni and Co become less siderophile with increasing pressure, and are equally siderophile at ~28 GPa, corresponding to a depth of 750 km in the present Earth. This in turn suggests (Wood 2011) that the iron in impacting asteroids, in which the cores had formed at low pressure, did not sink directly to the Earth's core, but spent time at an intermediate depth where it had time to re-equilibrate (i.e., reach chemical equilibrium) with the silicate mantle at high pressure before continuing to sink. One possible model that would be consistent with this result is a 750-km deep outer mantle kept molten by impact heating and radiogenic heat from ^{235}U, ^{238}U, ^{232}Th, ^{40}K (see Chap. 6) and short-lived radioactive isotopes such as ^{26}Al ($t_{1/2} = 7.17 \times 10^5$ year). The molten outer mantle would be separated from the molten iron core by a solid inner mantle. An incoming asteroid would melt on impact, and its iron core would sink in droplets through the molten outer mantle and pool at the top of the solid inner mantle, where it would have time to equilibrate with the silicate. When enough of the denser iron had accumulated above the lighter rock in the solid inner mantle, the system would become unstable and the iron would sink as large diapirs through the rock and into the core.

5.4.4.1 The Size of the Earth's Core

The fractions of the Earth taken up by the inner and outer core can be summarized as follows:

	Outer core (%)	Inner core (%)
By volume	15.7	0.7
By mass	30.8	1.7
By number of atoms	15.0	0.8

5.4.4.2 The Molten Outer Core

The state of this core must be molten.

The evidence includes the following:

1. The existence of the Earth's magnetic field, as described above.
2. S-waves do not propagate through the outer core, and are not seen.
3. Results from analysis of free oscillations of the Earth as a whole.
4. The character of the Earth's nutation (small wobbles in the direction of the Earth's spin axis): if you spin an egg, it behaves differently depending on whether it is hard-boiled (solid) or raw (fluid inside the shell); similarly, the spinning Earth responds differently to gravitational tugs depending on whether it is entirely solid or has a large molten core.
5. Convective and cyclonic motions occur in the fluid outer core, producing the Earth's magnetic field through dynamo action. Speeds appear to be about 10 km/year.

5.4.4.3 Non-Iron Composition of the Core

The observed density of the outer core is 10 ± 2 % less than that expected for pure iron at the outer core's temperature and pressure, and the inner core is 2.5 % less dense than pure iron at the inner core's temperature and pressure. At least one lighter element (possibly more) is therefore mixed in as an impurity. The element or elements involved are not known. Those that have been suggested (in the form of pro and con arguments) are:

1. Sulfur (S)

For:

- S alloys easily with iron, forming sulphides, at both low and high pressures.
- S lowers the melting temperature of the iron alloy compared to pure iron. This favors the formation of the observed solid inner core because, as the core temperature decreases over geological time, pure iron can separate out, solidify, and sink, forming the inner core, while the alloy in the outer core remains molten (Sect. 5.4.4.5).
- Iron sulphides are found in meteorites (0.7 ± 0.5 % by weight of the iron).
- S is depleted in the mantle compared to cosmic abundances. Enhanced sulfur in the core would account for part of this depletion.

Against:

- S in the core cannot explain all of the sulfur depletion in the mantle, even if sulfur makes up the entire light-element content of the core. At least some sulfur therefore must have been lost to space as a volatile during the formation of the Earth. (The case for sulfur in the core would be strengthened if none could have been lost to space, because then the only place the "missing" mantle sulfur could have gone would be to the core.)
- If sulfur makes up a significant fraction of the light-element content of the core, then it would in fact be more abundant in the Earth as a whole than some elements which are less volatile (less easily lost to space) than sulfur.

2. Oxygen (O)

For:

- Is definitely present, because chemical reactions at the core–mantle boundary necessarily add oxygen to the core. This oxygen will be mixed through the core by convection. (The question is, is it present in significant amounts?)
- Is abundant in the Earth (58 ± 2 % of the mantle by number of atoms).
- FeO becomes metallic at the high pressures found in the Earth's core, and therefore alloys easily with the molten iron there.

Against:

- Does not easily alloy with iron at low pressures. It could therefore not have alloyed with iron until well after the core had begun to form, and perhaps not until after the accretion of the Earth was complete. This may leave chemical reactions at the core boundary as the major source of oxygen in the core, and this may not be significant.

3. Silicon (Si) and Magnesium (Mg)

For:

- Necessarily added by chemical reactions at the core–mantle boundary, as with oxygen.
- Si and Mg relatively abundant (18 ± 7 % of the mantle by number of atoms).

Against:

- Smaller amounts available compared to oxygen.
- Si and Mg appear to be less reactive with molten iron than oxygen.

4. Hydrogen (H):

For:

- H is abundant in the universe, and also possibly at the Earth's surface during accretion and core formation.
- H alloys with iron easily at high pressures.

Against:

- Most of the Earth's hydrogen was lost to space; the amount (if any) in the core cannot explain the observed deficit compared to cosmic.
- The presence of H in the core seems to require the Earth to remain cold during accretion and subsequent core formation, to prevent the hydrogen from being lost to space. This is contrary to expectation.

5. Nickel (Ni)

For:

- Ni is present in iron meteorites (8 ± 7 % by weight).
- Ni is depleted in the mantle by an amount which would be completely accounted for if the core contained about 4 % by weight of nickel.
- Ni alloys easily with iron at both low and high pressures.
- Ni lowers the melting point of the iron alloy compared to pure iron and so favors formation of a solid inner core (as noted above for sulfur).

Against:

- Ni has almost the same density as iron, so it does not affect the need for light elements discussed above. In fact, there is *no* geophysical observation that directly indicates the presence of nickel in the core (or any that excludes it).

Clearly there are many candidates, some more likely than others.

5.4.4.4 The Inner Core

The inner core is apparently solid. There are two main lines of evidence:

1. S-waves appear to propagate through the inner core; this is not possible in a liquid.
2. P-wave velocities in the inner core are systematically higher parallel to the Earth's rotational axis (the polar or N–S direction) than parallel to the plane of the equator (Calvet et al. 2006 and references therein). A similar anisotropy is found in the uppermost part of the mantle and is caused by olivine crystals having a preferred direction of orientation (produced by shear associated with convection and plate tectonics). The same mechanism could operate in the inner core if it is also tectonically active (solid but with plastic flow), because solid iron is expected to have a hexagonally close-packed (hcp) crystal structure at the inner core's temperature and pressure, and hcp iron (or ε-Fe) is known to have anisotropic elastic properties.

Because of the high thermal conductivity of solid iron, the inner core is expected to approach isothermality on a timescale of the order of 10^9 years (Williams et al. 1987 and references therein). An age for the inner core can be derived from the fact that it is believed to play a significant role in the production of the Earth's magnetic field, and paleomagnetic measurements show a strong terrestrial magnetic field for at least the past 2.5–3.5 billion years. Although this may be insufficient time to achieve full isothermality, measured data suggest a maximum temperature difference between the center and the inner core boundary of 300 K or less, or about 10 % or less of the temperature difference across the outer core.

5.4.4.5 Formation of the Inner Core

If two substances (e.g., iron and sulfur) are miscible in the liquid phase but, because they have different crystal structures, are immiscible in the solid phase, the result is a *binary eutectic system*. Figure 5.21 illustrates the behavior of such a system, consisting of components A and B. (In the case of iron and sulfur, the two components of the binary system are Fe and FeS.)

If we gradually cool a completely molten mixture that has the eutectic composition (see Fig. 5.21 for terminology), then both components remain molten until the eutectic is reached, and both components crystallize simultaneously until the mixture is completely solid. Such a composition is called a *eutectic mix*. In this case, the system has a well-defined freezing (or melting) temperature, the *eutectic temperature*. However, if the initial composition is to the left of the eutectic in Fig. 5.21 (an example is shown by the thick arrows), then when the melt reaches the liquidus, A begins crystallizing while B and the remainder of A remain molten. As A is removed from the melt, the melt gradually becomes richer in B, so as the temperature continues to decrease, the melt "slides" along the liquidus toward

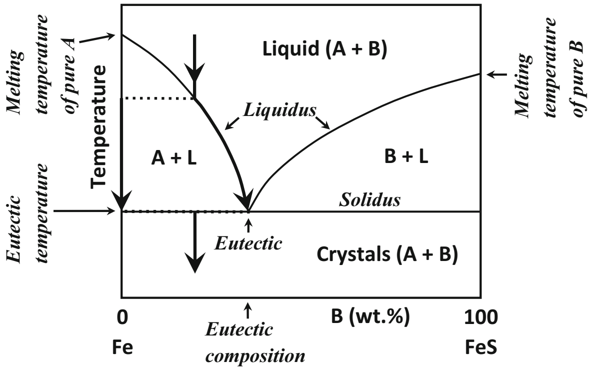


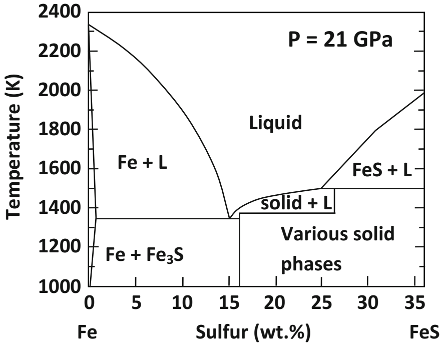
Fig. 5.21 Binary eutectic phase diagram for two components, A and B, that are miscible in the liquid phase and immiscible in the solid phase. The symbol L = liquid. The liquidus curves are the lines above which the entire system is liquid, and the solidus is the line below which the entire system is solid (a mixture of crystals of A and B). In the regions between the liquidus and the solidus, the mixture consists of crystals of one component within a liquid of both components. The eutectic is the point at which all three phases (A, B, and the melt) can exist simultaneously. The thick arrows show an example of cooling a melt that has an initial composition different from the eutectic composition

the eutectic as crystals of pure A continue to form. If the initial composition is to the right of the eutectic, then B crystallizes and the remaining melt slides to the left along the liquidus. In either case, when the melt reaches the eutectic, all of B and the remainder of A crystallize simultaneously until the mixture is completely solid. Eutectic mixes are considered in the contexts of the Moon's and Mercury's interiors in Sects. 8.5.3 and 9.1.4, respectively, and of Europa's and Titan's compositions in Milone and Wilson (2014), Sects. 13.1.2.2 and 13.1.3.1, respectively.

We now need to consider two different cases:

1. *Equilibrium crystallization*: If the crystals remain suspended (e.g., if the liquid is turbulently convective), then the total composition of crystals + melt remains constant. When the melt reaches the eutectic, both B and the remainder of A crystallize simultaneously in a ratio equal to the eutectic composition, but the resulting solid (a mixture of all crystals) has the same composition as the original melt.
2. *Fractional crystallization*: If the crystals sink to form a cumulate pile at the base of the melt and/or rise to form a crust, then they are removed from the melt. The final melt then crystallizes to form a solid with the eutectic composition. The result is that the originally-molten region has become differentiated into (depending on circumstances) perhaps three regions of different composition: a cumulate pile, a solidified final melt with the eutectic composition, and a crust. Depending on the initial melt and the minerals precipitating from it, the cumulate pile and crust can each be chemically differentiated as well (see Sect. 8.6 for an example).

Fig. 5.22 Binary eutectic phase diagram for Fe-S at 21 GPa pressure. The solidus is complicated by the existence of different compounds of Fe + S, including Fe_3S (shown) and Fe_2S. After Fei et al. (2000) with permission from the author and the Mineralogical Society of America. The symbol L = liquid



In the molten iron core of a terrestrial planet, if we take sulfur as the impurity, we can define Fe as component A and FeS as component B. The fraction of FeS (or S) is expected to be below the eutectic composition, so as the core cools, crystals of pure iron begin to form when the system reaches the liquidus (Fig. 5.22). Iron is denser than the melt, so the crystals sink, creating a solid inner core of pure iron surrounded by a molten Fe-FeS outer core. As the planet cools, the solid inner core grows as iron continues to precipitate, and the molten outer core becomes increasingly enriched in sulfur. When the system reaches the eutectic, both Fe and Fe_3S precipitate until the mixture is entirely solid.

Figure 5.23, adapted from Chudinovskikh and Boehler (2007), shows the eutectic composition of the Fe-S system as a function of pressure to 44 GPa. The break in the curve appears to arise because Fe and FeS form a simple binary eutectic system up to 10 GPa (Kamada et al. 2010), but a binary system with intermediate compounds at higher pressures: Fe_3S_2 above 14 GPa and Fe_2S and Fe_3S above 21 GPa. The eutectic composition appears to approach 10 wt% S (*read:* 10 % S by weight) asymptotically as pressure increases above ~40 GPa, which provides an upper limit to the sulfur content of the outer core if the precipitate is to be iron.

The fact that the inner core is 2.5 % less dense than pure iron under inner-core conditions requires that a light element be present in the inner core also. If the light element is sulfur, the required amount is 2.2–6.2 at.% (*read:* 2.2–6.2 % by number of atoms). The hcp phase of iron (ε-Fe) can contain >7.5 at.% (or 4.4 wt.%) S under inner-core conditions, so ε-Fe may be the only iron phase present; or, the inner core could contain a small fraction of Fe_3S (Kamada et al. 2010).

The melting temperature, T_M, of the core material as a function of pressure (and therefore of radius in the Earth) provides a lower limit to the geotherm, T_r, in the molten, outer core and an upper limit in the solid, inner core, and equals T_r at the inner core–outer core boundary (IOB). Thus, T_M provides an important constraint on the geotherm. However, experimental and theoretical uncertainties make T_M uncertain, as do uncertainties in the composition of the core material (Sect. 5.4.4.3); see Boehler (2000) for a review.

Static measurements of T_M in laser-heated diamond-anvil cells have been conducted for pure Fe up to pressures as high as 2 Mbar and in shock-heated

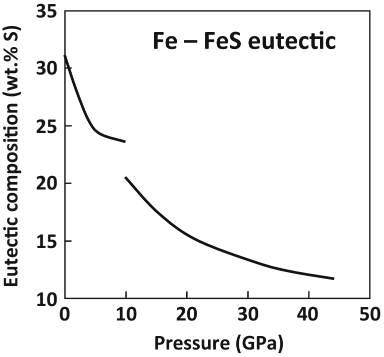


Fig. 5.23 Measured eutectic composition *vs.* pressure for Fe-FeS. See text for details. Adapted with permission from Chudinovskikh and Boehler (2007), and Elsevier, publisher of Earth and Planetary Science Letters

diamond-anvil cells up to about 4 Mbar, and to lower pressures for Fe–FeS systems including Fe_2S and Fe_3S, and for Fe–FeO systems. For comparison, the pressure is about 1.36 Mbar at the core-mantle boundary (CMB) and 3.30 Mbar at the IOB. Shock heating measurements are generally considered less accurate than static measurements because of various uncertainties, including the optical and thermal behaviors of the window material during the shock (Boehler 2000), but provide important constraints on the melting curves.

Several static diamond-anvil studies (Boehler 2000 and references therein) agree on T_M for pure iron in the range 3,100–3,400 K at the CMB, and, by extrapolation to higher pressures, 4,200–5,000 K at the IOB (Fig. 5 of Boehler 2000). As noted above, the presence of S or both S and O can lower T_M by a few hundred K. (Earlier studies indicated that FeO had a higher T_M than pure Fe, but more recent results suggest that, at core pressures, T_M is essentially the same for both FeO and Fe.) With T_M below 5,000 K at the IOB, and assuming temperature changes adiabatically throughout the outer core, T_r on the core side of the CMB is below 4,000 K.

The P- and S-wave speeds and the density as a function of radius are shown for the mantle in Fig. 5.24 and for the entire planet in Fig. 5.25. Finally, the march of pressure, gravitational acceleration, and density through the Earth for the preliminary reference model are plotted in Fig. 5.26.

Some recent analyses of deep-Earth seismic data (Ishii and Dziewonski 2002, 2003; Beghein and Trampert 2003; Calvet et al. 2006; Sun and Song 2008) suggest that the anisotropy of the inner core displays a marked change in character in its innermost part (r < 300–450 km, or ~590 km (Sun and Song 2008)). Although several different studies agree on the existence of this innermost inner core, its character remains uncertain. Ishii and Dziewonski (2003), from inverting travel-time data for P-waves and using a ray approximation, find that the fast axis in the innermost inner core could be tilted as much as 55° from the Earth's rotation axis. Even assuming that the fast axis is parallel to the Earth's rotation axis, as it is in the outer part of the inner core, they still find a distinct innermost inner core in that the anisotropy is noticeably stronger. Whether the anisotropy change occurs sharply at a boundary or gradually over a range of radii is unclear. Beghein and Trampert (2003) find the fast axis in the Earth's equatorial plane for P-waves and parallel to the Earth's

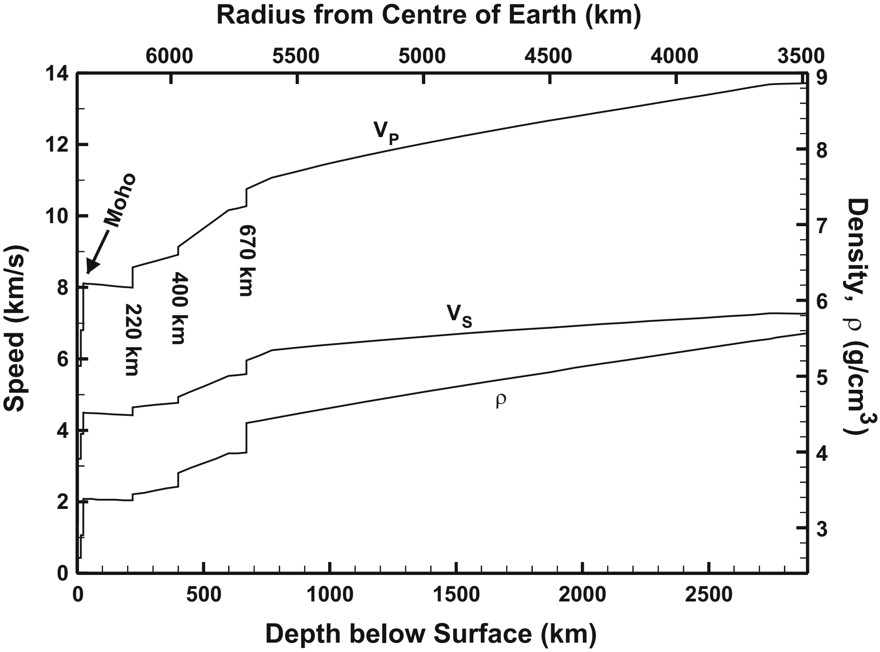


Fig. 5.24 The variation of P- and S-waves and the density through the crust and the mantle. Note the trend for all three quantities to increase with depth. Produced from data published by Dziewonski and Anderson (1981), Table II

rotation axis for S-waves. Calvet et al. (2006) perform an extensive analysis of P-wave data using a more thorough method than previous investigators. Their results agree with previous models for the outer part of the inner core, but for the innermost inner core the data are consistent with any of three basic models: (1) a weak anisotropy with the fast axis in the equatorial plane; (2) near isotropy; and (3) a strong anisotropy with the fast axis parallel to the Earth's rotation axis. In all three cases, the innermost inner core appears distinct in its properties from the outer part of the inner core, despite the lack of constraint on these properties.

This innermost inner core (IMIC) could arise through one of at least three scenarios (Ishii and Dziewonski 2002):

1. The IMIC could be a fossil from an early period of rapid differentiation of the Earth, with the rest of the inner core forming slowly later; in this case, the IMIC could be chemically distinct from the rest of the inner core.
2. The inner core affects the flow pattern in the outer core, and this flow pattern could have changed character when the inner core reached a certain size, altering the anisotropy of the iron core thereafter.
3. The different anisotropy could indicate a different phase of iron at the pressure and temperature of the IMIC.

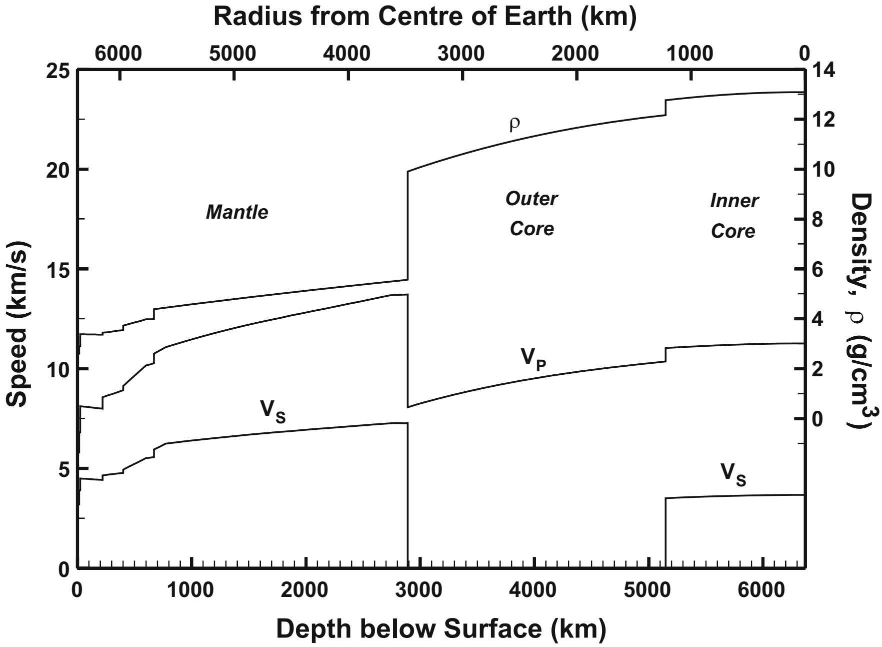


Fig. 5.25 As per Figure 5.24 but for the entire planet. Note the sudden decrease in both P and S velocities at the boundary between the core and the mantle, and the absence of S-waves in the outer core. Produced from data in Dziewonski and Anderson (1981, Table II)

After we discuss the heat flow and internal temperatures (Chap. 6) and the nature of the material in the interior of the Earth (Chap. 7), we will revisit the structure of the Earth and, later, the other terrestrial planets.

5.5 Planetary Surfaces

5.5.1 Impacts

The properties of the surfaces of the rocks reveal something of the history of the planet, especially when that surface has been unmodified for eons—like the surfaces of the Moon or Mercury. The extensive cratering on these surfaces shows the effect of a long period of intense bombardment, which the Earth, Venus, and Mars certainly did not avoid, and may have been more intense for them because of their greater gravitational attraction. The population of small colliders that we know as meteoroids ablates in the atmospheres of the Earth or Venus, but much larger objects—meters or larger in diameter—explosively dissipate much or all of their

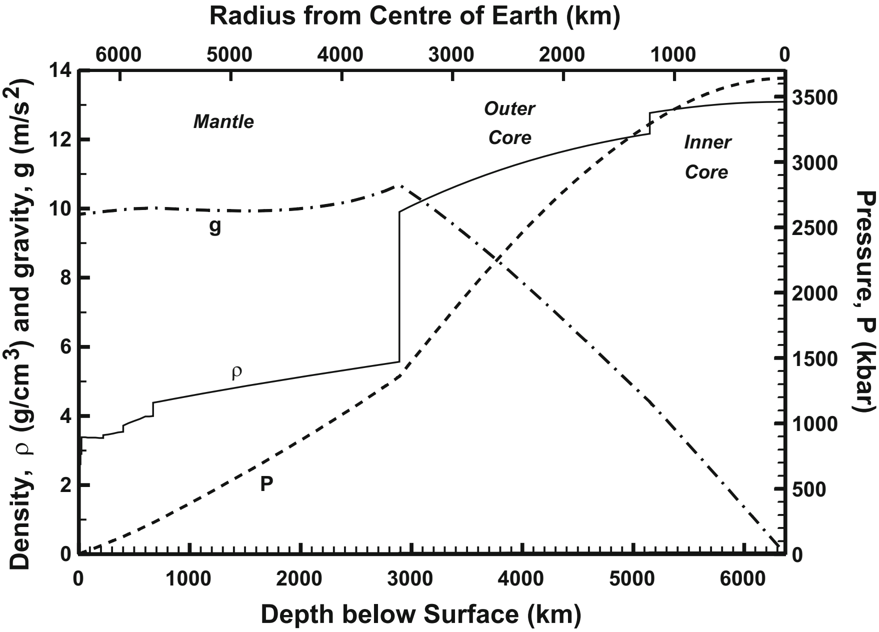


Fig. 5.26 The variation of density, pressure, and gravitational acceleration with radius and depth, according to the preliminary reference Earth model. Produced from data in Dziewonski and Anderson (1981, p. 27)

material. The energy per unit mass arises ultimately from the gravitational potential of the target body but more directly from the relative speed at impact:

$$E = v^2/2 \tag{5.38}$$

The surfaces of Venus, Earth, and Mars have shown significant modification since a period of intense bombardment in the solar system, some 4 Gy ago.

In the case of Mars, as we shall see, the modification has been selective; on Earth and Venus it has been extensive.

The effect of an impact depends critically on the velocity of the impact and the mass of the impactor. The mass of a meteoroid is unknown typically, but if its size can be determined, an estimate of its density leads to a mass. The density ranges from 1,000 kg/m^3 for a "rubble pile" asteroid (an aggregation of loosely packed material) to solid nickel–iron, ~8,000 kg/m^3. The speed of impact depends on the orbit of the impactor. If a meteoroid is initially traveling in a parallel path to the planet and has in effect no net speed with respect to it, the meteoroid will fall to the planet's surface with the escape velocity of that planet:

$$v_{\mathrm{esc}} = \left[2GM_{\mathrm{p}}/R_{\mathrm{p}}\right]^{\frac{1}{2}} \tag{5.39}$$

This quantity is 11.2 km/s for the Earth and only 5.01 km/s for Mars. A meteoroid of asteroidal origin is likely to have originated in the asteroid belt between Mars and Jupiter, although there is a relatively small population of objects even within the orbit of the Earth. There is also a considerable population of objects in the outer solar system, the "trans-Neptunian objects," which are more or less coplanar to the planets and travel in CCW orbits. These include the icy bodies of the Edgeworth-Kuiper Belt. Finally, far beyond the 100 or so au of this region is the spherically distributed Oort Cloud, from which we get the long-period comets. A cometary object with a semi-major axis, *a,* of 10,000 au and perihelion distance, *q*, of 1 au has an eccentricity (cf. Sects. 2.2 and 3.2):

$$e = 1 - q/a = 1 - 10^{-4} \approx 1 \tag{5.40}$$

The speed of an Oort cloud comet at perihelion is therefore just short of the escape velocity from 1 au. This escape speed is:

$$v_{esc} = [2GM_{\odot}/q]^{\frac{1}{2}} = 4.21 \times 10^{4}\mathrm{m/s} = 42.1\ \ \mathrm{km/s} \tag{5.41}$$

It is instructive to compute the speed of impact and the energy released if such a comet with perihelion at 1 au were to encounter the Earth. The important quantity to compute first is the speed of the comet relative to the Earth just before impact. This speed is greatest if the comet is in a retrograde orbit and encounters the Earth in a head-on collision, and least if the comet is in a prograde orbit and catches up to the Earth from behind.

The impact speed may be computed in two steps: (1) find the relative speed, v_1, of the comet and Earth if the Earth did not attract the comet gravitationally; and (2) use conservation of mechanical energy to compute the actual speed, v_2, after the comet falls from infinity (with initial speed v_1) to the surface of the Earth at $\mathrm{r} = \mathrm{R}_{\mathrm{E}}$.

Step 1: This is a relative motion problem. The perihelion speed of the comet is 42.1 km/s and the Earth's orbital speed (assuming a circular orbit) is 29.8 km/s. If the comet and the Earth are meeting head-on then the relative speed is $\mathrm{v}_{1,\mathrm{max}}$ = 42.1 km/s + 29.8 km/s = 71.9 km/s, and if the comet catches up from behind then $\mathrm{v}_{1,\mathrm{min}}$ = 42.1 km/s—29.8 km/s = 12.3 km/s.

Step 2: Conservation of mechanical energy:

$$K + U = \text{constant, where } K = \frac{1}{2}m_{comet}v^2 \text{ and } U = -\frac{GM_{Earth}m_{comet}}{r}$$

$\mathrm{U} = 0$ at $\mathrm{r} = \infty$, so

$$\frac{1}{2}\,\mathrm{m}_{comet}v_2^2 \quad - \frac{GM_{Earth}m_{comet}}{R_E} = \frac{1}{2}m_{comet}v_1^2 - 0$$

Solving for v_2,

$$v_2^2 \quad = v_1^2 + \frac{2GM_{Earth}}{R_E}$$

where, by (5.41), the second term on the right is the square of the Earth's escape speed. Then

$$v_{2,\min} = \sqrt{v_{1,\min}^2 + v_{esc}^2} = \sqrt{(12.3\ \ \mathrm{km/s})^2 + (11.2\ \ \mathrm{km/s})^2} = 16.6\ \ \mathrm{km/s}$$
$$v_{2,\max} = \sqrt{v_{1,\max}^2 + v_{esc}^2} = \sqrt{(71.9\ \ \mathrm{km/s})^2 + (11.2\ \ \mathrm{km/s})^2} = 72.8\ \ \mathrm{km/s}$$

The energy per unit mass involved in such a collision would be, from (5.38),

$$E_{\min} = \frac{1}{2}v_{2,\min}^2 = \frac{1}{2}\left(1.66 \times 10^4\ \mathrm{m/s}\right)^2 = 1.38 \times 10^8\ \mathrm{J/kg}$$

$$E_{\max} = \frac{1}{2}v_{2,\max}^2 = \frac{1}{2}\left(7.28 \times 10^4\ \mathrm{m/s}\right)^2 = 2.65 \times 10^9\ \mathrm{J/kg}$$

for Earth impact. Because the chemical energy released in a TNT explosion is 4.2×10^6 J/kg, the maximum energy released in a head-on cometary collision would be equivalent to ~630 kg of TNT per kg of impactor mass. Small impactors (meters across or less) ablate as they are passing through a planetary atmosphere, and fragments fall to the surface at the terminal velocity, not the escape velocity. Large objects (hundreds of meters to tens of kms), on the other hand, will not be slowed down very much and will impact with great violence, resulting in a very large crater and substantial deposition of material from the conical sheet that is excavated by the impact over the rim and at the center. The material at the forward edge and at ground zero will be vaporized by the high temperatures resulting from the transfer of kinetic energy; the high vapor pressure will then cause a recoil on the trailing edge of the material, which may break up and become distributed over a wide area or even be re-ejected.

In a planetary atmosphere, the material may be carried around the planet, eventually falling to the surface.

O'Keefe and Ahrens (1986) demonstrated that oblique impacts across a limited range of impact angles and speeds would be able to cause ejection of lightly shocked surface material through entrainment of the ejecta plume. Their Fig. 3 indicates that fragments between 10 cm and 100 cm can be ejected by a low-density impactor with a diameter between 1 and 100 m. This can account, e.g., for the SNC meteorites (Milone and Wilson 2014, Sect. 15.3.2). The relevant equation is

$$r_p = \frac{3\rho_g x}{2\rho_p} \tag{5.42}$$

where r_p and ρ_p are the plume-entrained fragment radius and density, respectively, ρ_g is the density of the high-velocity plume, and x is a plume radius.

A recent event demonstrated that meteoritic impacts are not merely of academic interest, but a clear and present danger! On February 15, 2013, an airburst near the Russian city of Chelyabinsk (about 300 km west of Tunguska, site of an even larger airburst in 1908) blew the glass out of windows and damaged many buildings, injuring 1,500 people. The meteoroid had entered the atmosphere over a region bordering four countries: Mongolia, China, Russia and Kazakhstan, at an angle of approach of ~16° and proceeding WNW, disintegrated at an altitude of ~23 km. From the brightness of the fireball (V = –28, 3× brighter than the Sun) the estimated diameter of the impacting object was ~ 17 m, its mass, ~ 1.1×10^7 kg, and the energy of the explosion equivalent to ~440 kt or 4.4×10^8 kg of TNT (Durda 2013). The Tunguska event measured 3–5 megatons.

We will discuss the roles played by impacts on the surfaces of the planets and moons of the solar system in later chapters (Chap. 8 for the Moon, Chap. 9 for the surfaces of the individual terrestrial planets, and Milone and Wilson (2014, Chap. 13) for the other moons of the solar system).

5.5.2 *Observing Planetary Surfaces*

5.5.2.1 Phase and Visibility

The *phase, q,* of a planet is the fraction of a planet's diameter that appears illuminated by the Sun as viewed from the Earth (or any other platform from which you are viewing the planet!—for now we assume that the Earth is the only viewing site).

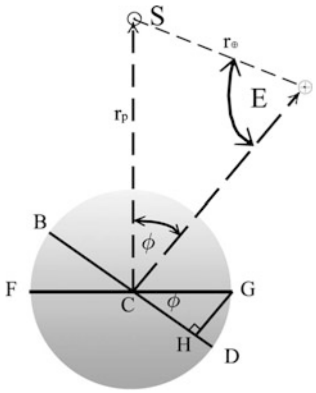
The *phase angle, ϕ,* is the angle at the center of the planet between the directions to the Sun and to the Earth. The relation between q and ϕ is:

$$q = \frac{1}{2}(1 + \cos\phi) \tag{5.43}$$

See Fig. 5.27 for the geometry needed to derive this equation.

Note that when $\phi = 0°$, the planet is fully illuminated (this is a possible configuration for all planets but for practical reasons only an exterior planet can be well viewed when fully illuminated, which for those planets occurs at opposition).

Fig. 5.27 Definition of phase angle



When $\phi = 90°$, half the planet appears illuminated. At a phase angle of 180°, $q = 0$. This is the case only at inferior conjunction, and is possible only for an interior planet.[6]

Although all phase angles are possible for an interior planet (whether easily viewable or not!), this is not true for exterior planets. In those cases, $0° \leq p < 90°$, but even sharper constraints can be found, so that $\mathrm{p}_{\max}$ and q_{min} may be specified for a given planet.

The sine law for plane triangles allows the phase angle of a planet to be calculated from the *elongation, E* (the angle between the Sun and the planet measured at the Earth, $\sphericalangle S \oplus C$ in Fig. 2.4), and the distances of the Earth from the Sun, r_0, and of the Sun from the planet, r_p, respectively, at any instant.

$$\sin\phi / r_\oplus = \sin E / r_p \tag{5.44}$$

From (5.44), for given values of $r_\oplus$ and r_p, the maximum phase angle, $\phi_{\max}$ (and therefore minimum phase, $q_{\min}$) of a superior planet occur at *quadrature* ($E = 90°$);

$$\sin\phi_{\max} = r_\oplus / r_p \tag{5.45}$$

From (5.44) and (5.45), for superior planets, $\phi_{\max} \to 90°$ only if $r_\mathrm{p} \to r_\oplus$. Mars has the smallest orbit beyond the Earth's and therefore attains the maximum departure from full phase that we see among the superior planets. The maximum possible value of $\phi_{\max}$ for Mars occurs on the extremely rare occasion when the

[6] *Interior*, that is, to the Earth's orbit. Classically, Mercury and Venus are "inferior planets," because they orbit below the orbit of the Sun in the Ptolemaic, geocentric universe. Similarly, planets *exterior* to the Earth's orbit, Mars on out, were classically referred to as "superior planets," because their orbits lay above that of the Sun. In modern usage, "inferior" and "superior" have acquired heliocentric meanings and are used interchangeably with "interior" and "exterior" respectively. See Chapter 1 and Kelley and Milone (2011) for details, configurations, and terminology.

Earth is at aphelion at the same instant that Mars at *quadrature* ($E = 90°$) is at perihelion. From (5.45),

$$\phi_{\max} = \arcsin(1.017/1.381) = 47\overset{\circ}{.}4$$

which, when inserted in (5.43), leads to a minimum possible phase, $q_{\min} = 0.838$. Because of the low probability of the required conditions occurring, observed phases of Mars will almost always be greater than this.

5.6 Addendum: Properties of Legendre Polynomials and Associated Legendre Functions

In simpler form, the function of interest in (5.7), namely,

$$\begin{aligned}\sum_{n=0}^{\infty} P_n(\cos\theta)\left(\frac{a}{r}\right)^n &= \left[1 - 2(\cos\theta)\left(\frac{a}{r}\right) + \left(\frac{a}{r}\right)^2\right]^{-1/2} \\ &= \sum_{n=0}^{\infty} \frac{(2n)!}{2^{2n}(n!)^2}\left[2(\cos\theta)\left(\frac{a}{r}\right) - \left(\frac{a}{r}\right)^2\right]^n\end{aligned} \tag{5.46}$$

is convergent for $(a/r) < 1$. The *Legendre polynomial* part of this function, $P_n(\cos\theta)$, is written as;

$$P_n(x) = \sum_{k=0}^{[n/2]} (-1)^k \frac{(2n-2k)!}{2^n k!(n-k)!(n-2k)!} x^{n-2k} \tag{5.47}$$

where $x \equiv \cos\theta$, and the upper limit on the summation $[n/2] = n/2$ for even, and (n − 1)/2 for odd, n. Whence, $P_0(x) = 1, P_1(x) = x, P_2(x) = \frac{3}{2}x^2 - \frac{1}{2} = \frac{1}{2}(3x^2 - 1)$, $P_3(x) = \frac{1}{2}(5x^3 - 3x)$, and, generally,

$(2n + 1)xP_n(x) = (n + 1)P_{n+1}(x) + nP_{n-1}(x)$, $n = 1, 2, 3, \ldots$, Also $\partial P_{n+1}/\partial x = (n + 1)P_n(x) + x\partial P_n(x)/\partial x.$

Further, we can write,

$$P_n(x) = \frac{1}{2^n n!} = \left(\frac{d^n}{dx^n}\right)\left(x^2 - 1\right)^n \tag{5.48}$$

Finally, where "m = 0/1" designates "either 0 or 1," $P_n(x)$ can be written

$$P_n(\cos\theta) = \sum_{m=0/1}^{n} a_m \cos m\theta = \sum_{m=0/1}^{n} a_m\left(\mathrm{e}^{\mathrm{i}m\theta} + \mathrm{e}^{-\mathrm{i}m\theta}\right) \tag{5.49}$$

The *Associated Legendre functions* are characterized by:

$$P_n^m(x) = \left(1 - x^2\right)^{m/2}\left(\frac{\mathrm{d}^m}{\mathrm{d}x^m}\right) P_n(x) \tag{5.50}$$

where $m \leq n$. Thus, $\mathrm{P}_1{}^1(x) = (1 - x^2)^{1/2} = \sin\theta$; $P_2{}^1(x) = 3x(1 - x^2)^{1/2} = 3\cos\theta$ $\sin\theta$, $\mathrm{P}_2{}^2(x) = 3\sin^2\theta, \ldots.$

Challenges

[5.1] Derive equation (5.3b) in footnote 3.

[5.2] Assume that a meteoroid originates from a direct orbit with $a = 1.333$ au and $e = 0.25$, and collides with the Earth. Compute its speed at 1 au and on impact with Earth. If the object has a diameter of 100 m and a mean density of 3,300 kg/m^3, compute (a) the energy per unit mass and (b) the total energy of the explosion. (c) What do you suppose will happen if the object strikes an ocean rather than land? Will it leave a crater, for instance?

[5.3] Derive equation (5.34) and demonstrate why an exterior planet cannot be observed at phase angles $\leq 90°$. Show also that the maximum observed phase angle for an exterior planet occurs when the planet is at quadrature.

[5.4] Prove that the phase is a minimum when the phase angle is a maximum. Compute the minimum observable phases for the asteroid Ceres, Jupiter, and Pluto. (Orbital data are available in Milone and Wilson (2014), Chaps. 15, 12, and 13, respectively.)

[5.5] From the data and relations given in Sect. 5.5.1, calculate the speed at which the Chelyabinsk meteor exploded.

[5.6] Derive the equation of hydrostatic equilibrium (5.17), from the basic definition of the pressure on a cylinder of gas of cross-sectional area A and height h, subject to a gravitational acceleration g in a gas of density ρ.

References

Beghein, C., Trampert, J.: Robust normal mode constraints on inner-core anisotropy from model space search. Science **299**, 552–555 (2003)

Boehler, R.: High-pressure experiments and the phase diagram of lower mantle and core materials. Rev. Geophys. **38**, 221–245 (2000)

Calvet, M., Chevrot, S., Souriau, A.: P-wave propagation in transversely isotropic media II. Application to inner core anisotropy: effects of data averaging, parametrization and a priori information. Phys. Earth Planet. Inter. **156**, 21–40 (2006)

Chudinovskikh, L., Boehler, R.: Eutectic melting in the system Fe–S to 44 GPa. Earth Planet. Sci. Lett. **257**, 97–103 (2007)

Consolmagno, G.J., Schaefer, M.W.: *Worlds Apart: A Textbook in Planetary Sciences*. Prentice Hall, London (1994)

Deschamps, F., Kaminski, E., Tackley, P.: A deep mantle origin for the primitive signature of ocean island basalt. Nat. Geosci. **4**, 879–882 (2011)

Durda, D.: The Chelyabinsk super-meteor. Sky Telesc **125**(June), 24–31 (2013)

Dziewonski, A.M., Anderson, D.L.: Preliminary reference earth model. Phys. Earth Planet. Inter. **25**, 297–356 (1981)

Fei, Y., Li, J., Bertka, C.M., Prewitt, C.T.: Structure type and bulk modulus of Fe3S, a new iron-sulfur compound. Am. Mineral. **85**, 1830–1833 (2000)

Fortney, J.J.: The structure of Jupiter, Saturn, and exoplanets: key questions for high-pressure experiments. Astrophys Space Sci **307**, 279–283 (2007)

Gutenberg, B; Richter, C. F.: Magnitude and energy of earthquakes. Science **83** (2147), 183–185 (1936)

Henning, W.G., O'Connell, R.J., Sasselov, D.D.: Tidally heated terrestrial planets: viscoelastic response models. Astrophys J **707**, 1000 (2009)

Hunter, G.W., Ponchak, G.E., Dyson, R.W., Beheim, G.M., Scardelletti, M.C., Meredith, R.D., Taylor, B., Kiefer, W. S. Development of a High Temperature Venus Seismometer and Extreme Environment Testing Chamber. International Workshop on Instrumentation for Planetary Missions, Greenbelt, Maryland. Abstract 1133 (2012)

Ishii, M., Dziewonski, A.: The innermost inner core of the earth: evidence for a change in anosotropic behavior at the radius of about 300 km. Proc. Natl. Acad. Sci. U. S. A. **99**, 14026–14030 (2002)

Ishii, M., Dziewonski, A.: Distinct seismic anisotropy at the centre of the earth. Phys. Earth Planet. Inter. **140**, 203–217 (2003)

Islam, M.A., Ebihara, M., Kojima, H. Chemical Compositions and Alteration of Primitive Carbonaceous Chondrites. 43rd Lunar and Planetary Science Conference, abstract 1974 (2012)

Kamada, S., Terasaki, H., Ohtani, E., Sakai, T., Kikegawa, T., Ohishi, Y., Hirao, N., Sata, N., Kondo, T.: Phase relationships of the Fe-FeS system in conditions up to the Earth's outer core. Earth Planet. Sci. Lett. **294**, 94–100 (2010)

Kelley, D.H., Milone, E.F.: *Exploring Ancient Skies*, 2nd edn. Springer, New York (2011)

Ksanfomaliti, L.V., Zubkova, V.M., Morozov, M.A., Petrova, E.V.: Microseisms at the *Venera 13* and *Venera 14* landing sites. Pis'ma Astronomicheskii Zhurnal **8**, 444–447 (1982)

Li, J., Agee, C.B.: Geochemistry of mantle-core differentiation at high pressure. Nature **381**, 686–689 (1996)

Lorenz, R.D., Nakamura, Y. Viking Seismometer Record: Data Restoration and Dust Devil Search. 44th Lunar and Planetary Sciences Conference, abstract 1178 (2013)

Love, A.E.H.: Some Problems of Geodynamics. Cambridge University Press, Cambridge (1911)

Milone, E.F., Wilson, W.J.F.: Solar System Astrophysics: Planetary Atmospheres and the Outer Solar System, 2nd edn. Springer, New York (2014)

O'Keefe, J.D., Ahrens, T.J.: Oblique impact: a process for obtaining meteorite samples from other planets. Science **234**, 346–349 (1986)

Richter, C. F.: An instrumental earthquake magnitude scale. Bul. Seismological Soc. of Amer. **25** (1), 1–32 (1935).

Stewart, A., Schmidt, M.W., van Westrenen, W., Liebske, C.: Mars: a New core-crystallization regime. Science **316**, 1323–1325 (2007)

Sun, X., Song, X.: The inner inner core of the earth: texturing of iron crystals from three-dimensional seismic anisotropy. Earth Planet. Sci. Lett. **269**, 56–65 (2008)

Tromp, J.: Inner-core anisotropy and rotation. Annu. Rev. Earth Planet. Sci. **29**, 47–69 (2001)

Williams, Q., Jeanloz, R., Bass, J., Svendsen, B., Ahrens, T.J.: The melting curve of iron to 250 GPa: a constraint on the temperature at the Earth's center. Science **236**, 181–182 (1987)

Wood, B.: The formation and differentiation of earth. Phys Today **64**, 40–45 (2011)

Chapter 6
Planetary Heat Flow and Temperatures

The surface temperatures of planets in our solar system currently depend basically on four quantities:

1. The luminosity of the Sun
2. The distance of the planet from the Sun
3. The planetary bolometric albedo
4. The heat welling up from the interior

The first two of these determine the solar energy flux reaching the planet (the planet's solar constant) and the third determines the energy flux actually absorbed by the planet. In the early solar system, energy from frequent large impacts was a major contribution to planetary heat budgets. At present, it is not (although some of the heat energy accumulated from that source is still retained in the deep interiors of all the planets), so only the above four factors will be discussed here. A planet's surface temperature is then determined by the equilibrium condition that the energy absorbed from the Sun and the energy welling up from the interior must together equal the energy radiated by the planet. The total emitted radiation of planets generally exceeds that absorbed from the Sun, some by significant amounts. Saturn, for example, radiates 78% more heat than it receives, the excess coming from internal heat sources. On the other hand, the internal heat sources in the terrestrial planets are far less important than solar radiation, as we demonstrate below for the Earth.

To make a long story short, the total power radiated by the interior sources is 4.7×10^{13} W (47 TW, where 1 TW $=$ 1 terawatt $= 10^{12}$ W), corresponding to an average heat flux from Earth's interior of 0.092 W/m^2. By comparison, the radiative power absorbed from the Sun is 1.09×10^{17} W.

First we consider the heat flow from the interior, and the resulting geotherms for the Earth. In later chapters we will deal with the other terrestrial planets in a comparative way and, in Milone and Wilson (2014) Sect. 12.6, with the gas giants.

E.F. Milone and W.J.F. Wilson, *Solar System Astrophysics: Background Science and the Inner Solar System*, Astronomy and Astrophysics Library,
DOI 10.1007/978-1-4614-8848-4_6, © Springer Science+Business Media New York 2014

6.1 Heat Flow

The Earth's luminosity, L, is the total heat escaping from the Earth per second at all wavelengths. A determination of L that makes use of a dataset of 38,347 heat flow measurements around the world supplemented by a half-space cooling (HSC) model for young oceanic crust (<64.5 Ma), yields a value

$$L = 47 \pm 2\,\mathrm{TW}$$

(Davies and Davies 2010). The treatment of young oceanic crust is important because it is the site of a significant fraction of the Earth's heat loss. In its simplest form, the HSC model treats the mantle and lithosphere as a semi-infinite slab with an upper boundary at constant deep-oceanic temperature, T_s, at the top of the lithosphere and a constant mantle temperature, T_m, at infinite depth. Where fresh magma reaches the surface along a mid-ocean ridge, temperature is taken as constant with depth, z; i.e., initially $T(z) = T_m$. Subsequently, the thermal conductivity, κ, of the crust determines how the temperature gradient and heat flow vary with age and therefore with position as the new oceanic crust is pushed away from the ridge.

With $L = 47$ TW, the global average heat flux from the interior (Fig. 6.1) is:

$$F_{AVG} = \frac{L}{4\pi R^2} = \frac{4.7 \times 10^{13}\ \mathrm{W}}{4\pi\left(6.371 \times 10^{6}\ \mathrm{m}\right)^2} = 0.092\ \mathrm{W/m^2} = 92\ \mathrm{mW/m^2}$$

6.1.1 Sources of Internal Heat

The sources of heat can be classified into two types:

1. *Primordial heat*. Heat produced during the formation of the Earth; the Earth is still cooling down as a result.
2. *Ongoing processes*. Heat being generated inside the Earth now. We summarize the evidence for each in point form.

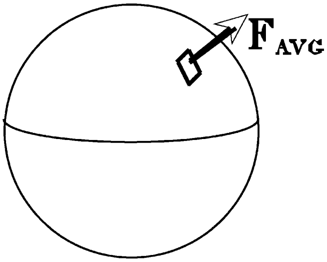


Fig. 6.1 Earth's heat flux

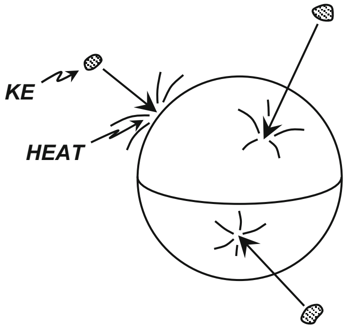


Fig. 6.2 Accretion heating

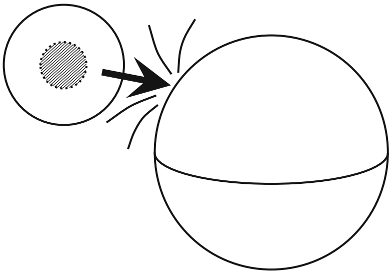


Fig. 6.3 Impact heating in a major event

6.1.1.1 Primordial Heat

Primordial heat has several likely sources, separately or in combination.

(1) Accretional Heat

Figure 6.2 illustrates impacts during early accretion. The source is the heat generated by the sudden stopping of incoming planetesimals during the accretion and later bombardment of the early Earth. Such events caused partial melting of the Earth.

(2) Possible Moon-Forming Impact

Figure 6.3 illustrates this major event (see Sect. 8.7). This event had three main features:

- ~Mars-sized impactor struck the Earth off-center.
- Core sank into the Earth's core.
- Mantle and part of the Earth's mantle were expelled to form the Moon.

(3) Short-Lived Radioactive Isotopes

- Radiogenic heating caused, for example, by $^{26}\mathrm{Al} \rightarrow {}^{26}\mathrm{Mg}$ (half-life $t_{1/2} = 7.4 \times 10^5$ y).

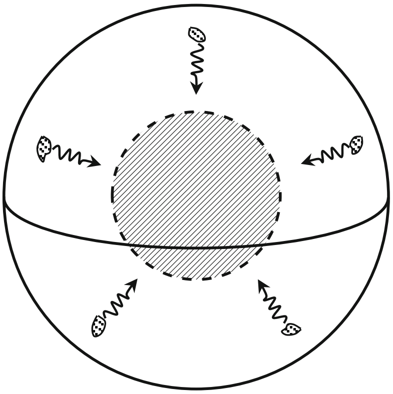


Fig. 6.4 Trickle-down heating from differentiation

- The energy released through the decay heats the Earth.
- With a half-life of 7.4 × 10^5 y, ^{26}Al is no longer left in the Earth. Most of the decay and heating took place during the first few half-lives (say 5 million years), so this is a component of the primordial heat.

(4) Differentiation of the Earth

Figure 6.4 illustrates the decrease in potential energy (PE) of droplets and larger blobs of liquid iron as they trickle down to create the core.

- PE is converted to kinetic energy (KE) in the fall toward the core.
- KE is converted into heat by friction and "collisions" with surrounding rock.
- The net result is that gravitational energy is released to form heat.
- The core was formed within about the first 40–100 million years (Wood 2011), therefore this source is included as a component of the primordial heat of the Earth.

(5) Present-Day Heat Flow due to Primordial Heat

As described below, on-going processes in the mantle account for only ~20 TW of the observed 47 TW heat flow through the Earth's surface. 5–15 TW also flows into the mantle from the core, an uncertain part of which is primordial, and the remaining ~12–22 TW (depending on the heat flow from the core) is primordial heat from the mantle. Because the primordial heat flow is not being replaced by on-going processes, it results in a secular cooling of the mantle.

6.1.1.2 Ongoing Processes

Figure 6.5 illustrates the Earth's interior heat budget (Lay et al. 2008; Arevalo et al. 2009).

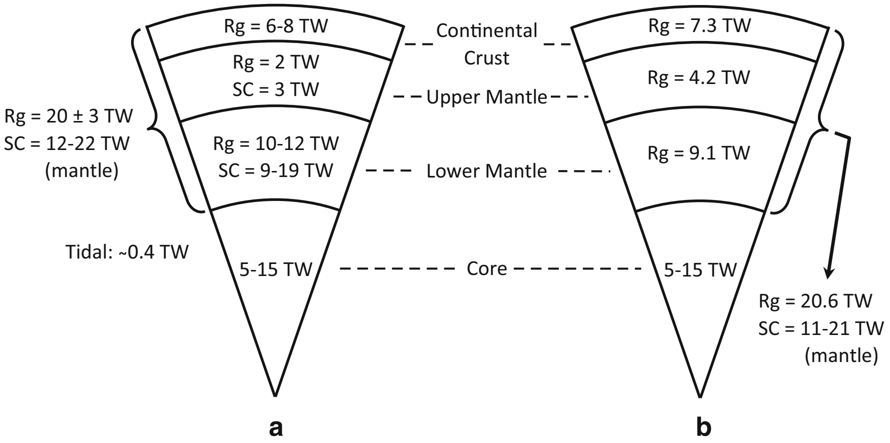


Fig. 6.5 Two recent models for interior heat flow in the Earth. (**a**) Data from Lay et al. (2008). (**b**) Data from Arevalo et al. (2009). Symbols: *Rg* radiogenic heat; *SC* secular cooling. Here, both models are adjusted to a total heat flow of 47 TW (Davies and Davies 2010)

The most important on-going source of heat in the Earth is radiogenic heat, accounting for ~43% of the 47 TW reaching the Earth's surface. It arises primarily from the following radioactive decay series[1]:

$$\begin{array}{ll} {}^{235}\mathrm{U} \rightarrow \ldots \rightarrow {}^{207}\mathrm{Pb} & t_{1/2} = 0.704 \times 10^9 \text{ y} \\ {}^{238}\mathrm{U} \rightarrow \ldots \rightarrow {}^{206}\mathrm{Pb} & t_{1/2} = 4.468 \times 10^9 \text{ y} \\ {}^{232}\mathrm{Th} \rightarrow \ldots \rightarrow {}^{208}\mathrm{Pb} & t_{1/2} = 14.05 \times 10^9 \text{ y} \\ {}^{40}\mathrm{K} \rightarrow {}^{40}\mathrm{A} \text{ or} {}^{40}\mathrm{Ca} & t_{1/2} = 1.248 \times 10^9 \text{ y} \end{array}$$

At the present time, ^{238}U and ^{232}Th contribute approximately 8 TW each and ^{40}K contributes about 4 TW (Fig. 8 of Arevalo et al. 2009). ^{235}U no longer contributes significantly because of its shorter half life, but it was an important contributor in the early Earth.

Independent support for the values of radiogenic heat shown in Fig. 6.5 is provided by geoneutrino measurements; i.e., neutrinos produced within the Earth by the decay series listed above. Although a longer time interval is required to reduce the uncertainty and the present detectors are not sensitive to the energy range

[1] For a single decay step, such as ^{14}C → ^{14}N, the half-life, $t_{1/2}$, is the time over which half of the parent nuclei (^{14}C) decay. In a decay series, each step of the series has a different half-life, but the rate of progress of the series is limited by the isotope with the longest half-life. In the case of the first three series, above, the isotope with the longest half-life is the first one in the series, so the three half-lives quoted are those of ^{235}U, ^{238}U, and ^{232}Th, respectively. Radioactive decay heats the Earth by releasing energetic α and β^- particles (^{4}He nuclei and electrons, respectively) that deposit their energy by collisions with atoms or ions in the surrounding rock. Radioactivity and its effects are discussed in more detail in Milone and Wilson (2014), Chap. 15.

of neutrinos produced by ^{40}K decay, measurements extending from 2002 to 2009 yielded a value of $20^{+8.8}_{-8.6}$ TW for radiogenic heat from ^{238}U and ^{232}Th (Gando et al 2011). This result is consistent within error of the 16 TW given above for these two decay chains.

The heat flux from the core is believed to arise from several causes:

(1) Latent heat is released by solidification and crystallization of the molten iron in the outer core, as it cools.
(2) Gravitational potential energy is converted to heat as the iron core shrinks as it cools, and also as iron crystals fall inward to form the solid, inner core. (The mantle also falls downward onto the shrinking iron core.)
(3) Radiogenesis may occur in the core, but little is as yet known. A high sulfur content could allow enough potassium to be present to generate as much as 2 TW (Lay et al 2008).
(4) Secular cooling.

Finally, there is a small contribution to interior heat generation by tidal effects. As discussed in Sect. 3.7.2, the Sun and Moon raise tidal bulges on the Earth, and the rotation of the Earth through these bulges creates torques that slow the Earth's rotation. As a result, the day is lengthening by 2.3 ms per century at the present time, but this rate varies with time: over the past 2,700 years the average rate has been 1.7 ms per century (Stephenson and Morrison 1995, pp. 188–193). The dominant torque is exerted by the lunar oceanic tides interacting with the continents, with smaller effects from solar oceanic tides and smaller still from solid-Earth tides.

The tidal bulge from the Moon also exerts a forward torque on the Moon that causes the Moon to recede from the Earth at a present rate of 3.8 cm/year. However, the Moon gains mechanical energy at a much smaller rate than the Earth loses mechanical energy, so most of the decrease in rotational kinetic energy of the Earth shows up as heat. Of this, only ~0.4 TW is produced by solid-Earth tides and therefore contributes to the interior heat flow (Lay et al. 2008, Fig. 1). It is interesting, nevertheless, to see what fraction this is of the total amount of tidal heating, and this can be estimated as follows. The rotational kinetic energy of an object of moment of inertia I rotating at angular speed $\omega = 2\pi/T$ (where $T = 1$ day $= 86{,}400$ s for the Earth) is

$$K = \frac{1}{2} I\omega^2 \tag{6.1}$$

If ω is changing, then the rate of change of the object's kinetic energy is

$$\frac{dK}{dt} = \frac{d}{dt}\left(\frac{1}{2} I\omega^2\right) = I\omega\frac{d\omega}{dt} = I\omega\alpha \tag{6.2}$$

where α is the angular acceleration,

$$\alpha \equiv \frac{d\omega}{dt} = \frac{d}{dt}\frac{2\pi}{T} = -\frac{2\pi}{T^2}\frac{dT}{dt} \tag{6.3}$$

The 2,700-year average value of *dT*/*dt* is 0.0017 s per century, so

$$\frac{dT}{dt} = \frac{0.0017 \text{ s}}{(100 \text{ y})(365.26 \text{ d/y})(86400 \text{ s/d})} = 5.39 \times 10^{-13}$$

The moment of inertia, *I*, of the Earth about its polar axis is in the approximate range 8.01 × 10^{37} kg · m^2 to 8.04 × 10^{37} kg · m^2. If we assume that the kinetic energy "lost" by the Earth is converted entirely into heat, then the rate of heat production, P_{tidal}, in the Earth by tidal action is the negative of the rate of change of the Earth's rotational kinetic energy:

$$\begin{aligned} P_{tidal} &= -\frac{dK}{dT} = I\omega\alpha = -I\left(\frac{2\pi}{T}\right)\left(-\frac{2\pi}{T^2}\frac{dT}{dt}\right) = \frac{4\pi^2 I}{T^3}\frac{dT}{dt} \\ &= \frac{4\pi^2\left(8.04 \times 10^{37} \text{ kg} \cdot \text{m}^2\right)}{(86400 \text{ s})^3}\left(5.39 \times 10^{-13}\right) = 2.65 \times 10^{12} \text{ W} = 2.65 \text{ TW} \end{aligned} \tag{6.4}$$

P_{tidal} is linear in *dT*/*dt*, so, considering that the present rate of increase in the length of the day is 2.3 ms/century, the present value of P_{tidal} is (2.3/1.7)(2.65 TW) = 3.59 TW.

Thus, ~10–15 % of the tidal heating occurs in the Earth's interior, and this in turn contributes <1 % of the Earth's interior heat budget.

6.1.2 Methods of Energy Transport

Energy can be transported in any of three ways: radiation, conduction, and convection.

6.1.2.1 Radiation

Long-wavelength radio waves (λ $\gtrsim$ 1 km) can propagate horizontally in the crust for thousands of kilometers, through low-conductivity rocks such as granites, gneisses and basalts between ~3 and 7 km depth. These rocks form a natural waveguide between the higher-conductivity materials above and below this depth, allowing long-distance, sub-surface radio communication. Low-conductivity layers at shallower depths can also allow localized radio communication over distances of several tens of kilometers. However, the Earth's crust is effectively opaque to shorter-wavelength electromagnetic radiation, and the outward transport of heat by the longer wavelengths is expected to be negligible compared to conduction and convection.

6.1.2.2 Conduction and Convection

Conduction involves the slow diffusion of heat from a warmer region to a cooler region, whereas convection involves the physical transport of the warmer material to a new site. Physical transport is faster than diffusion, so when convection occurs it is very efficient and by far dominates conduction.

A test for convection is provided by the *Rayleigh number*:

$$Ra = \frac{\text{buoyancy force } (\text{favoring convection})}{\text{viscous drag force (hindering convection)}}$$

The critical Rayleigh number for convection is $Ra_c \sim 2000$, where the subscript c stands for "critical." The condition is then,

$$(Ra > Ra_c \Rightarrow \text{convection})$$

The Rayleigh number for a fluid heated externally (e.g., from below) is given by

$$Ra = \frac{g\alpha\Delta T h^3}{\nu\kappa} \tag{6.5a}$$

and for a fluid heated internally, i.e., by heat sources within the fluid,

$$Ra = \frac{g\alpha q h^5}{\nu\kappa k} = \frac{g\alpha\rho q' h^5}{\nu\kappa k} \tag{6.5b}$$

(e.g., Stein et al. 2013, where the viscosity in their equations (4) and (5) is the dynamic viscosity, $\eta = \nu\rho$). Symbols and reference values for the terrestrial mantle (Stacey and Davis 2008; Schubert et al. 2001) are,

ρ = density ~4 g cm^{-3} = 4 × 10^3 kg m^{-3}
α = thermal expansion coefficient ~2 × 10^{-5} K^{-1}
g = acceleration due to gravity ~10 m s^{-2}
κ = thermal diffusivity ~10^{-6} m^2 s^{-1}
v = coefficient of kinematic viscosity ~2.5 × 10^{17} m^2 s^{-1}
h = thickness of convecting layer ~2,900 km = 2.9 × 10^6 m
k = thermal conductivity ~4.6 W m^{-1} K^{-1}
q' = rate of internal heat production per unit mass ~5.1 × 10^{-12} W kg^{-1}
$q = \rho q'$ = rate of internal heat production per unit volume (W m^{-3})
ΔT = temperature difference across convecting layer ~2,500 K

(6.5a) often referred to as the Bénard-Rayleigh number. Both external (5–15 TW from the core) and internal heating (~13 TW from radioactive isotopes) occur in the mantle (Fig. 6.5). The internal heating is strong enough that (6.5b) is generally used for mantle convection. From the reference values above, (6.5b) gives

$$Ra \sim 7 \times 10^8 >> \mathrm{Ra_c} \Rightarrow \text{Convective}$$

For comparison, (6.5a) yields $Ra \sim 5 \times 10^7$, which is also $\gg Ra_c$, Therefore, heat is transported out through the mantle by convection because warmer rock is buoyant and rises; cooler rock is denser and sinks. Under pressure, the rock in the mantle behaves like a high-viscosity fluid and undergoes plastic deformation (like the ice in a glacier). The speeds of convective motion are ~ cm per year.

Convection can be either laminar (smooth flow) or turbulent. A useful parameter to determine which applies is the *Reynold's number*:

$$Re = \frac{u\,d}{\nu} \tag{6.6}$$

where u is the speed of flow, d is the thickness of the convecting layer, and v is the kinematic viscosity. The flow is laminar if $Re \ll 1$ and turbulent if $Re \gg 1$. The best estimates for the Reynold's number in the mantle are around 10^{-21} (Anderson 1989, p. 255), in agreement with that calculated with the data above, so the flow is almost certainly laminar.

In the crust, heat transport is by conduction (e.g., Davies and Davies 2010), with a small fraction of the total being carried by advection in active volcanoes. In addition, in oceanic crust of ages up to several tens of millions of years, the upper several hundred meters of the crust is sufficiently permeable to seawater for the conductive heat flow arriving from below to be carried by hydrothermal circulation of seawater in these regions (Davis et al 1992).

6.1.2.3 Nature of the Convection

There are two major possibilities:

(1) Whole-Mantle Convection

Convection mixes the material; therefore, this model implies that the upper and lower mantles have basically the same composition, although segregation of rising rock into basaltic magma and olivine/pyroxene solids still takes place in the upper mantle.

With this model, the 400 and 650 km discontinuities must be due to phase transitions only, and not composition discontinuities.

Figure 6.6 illustrates this case. It also shows two features of the continental drift theory of modern geology: the uplift and spreading at a crustal plate interface and the subduction of a plate at another. Note that the descending portion of the crust, subducting under the thicker continental plate at the left, is

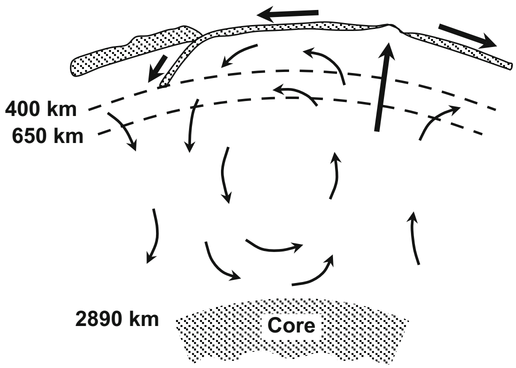


Fig. 6.6 Whole-mantle convection

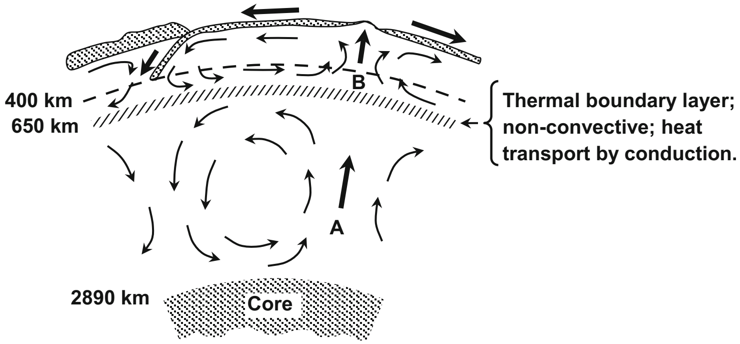

400 km
650 km
B
A
2890 km
Core
Thermal boundary layer; non-convective; heat transport by conduction.

Fig. 6.7 Two-layer mantle convection

free, in this model, to be reabsorbed fully into the deepest part of the upper mantle.

(2) Two-Layer Convection

In this case, illustrated in Fig. 6.7, the lower mantle and the upper mantle convect separately. The lower mantle convection drives the upper mantle convection, but the two are not *mechanically* coupled. There is a thermal boundary layer separating them at the 650 km discontinuity. This layer is non-convecting, with heat transport across it being by conduction.

Hot rock rising in the lower mantle creates a hot spot in the boundary layer. This in turn heats the rock at the base of the upper mantle, which then rises and drives the upper mantle convection. In this case, the 650 km discontinuity could also be a composition discontinuity, as well as a phase transition.

The lower mantle may be more silica-rich (that is, having more $MgSiO_3$ perovskite, with Mg:Si ~1:1) than the upper mantle (which has more Mg_2SiO_4 olivine, with Mg:Si ~2:1). The lower mantle may also be more iron-rich than

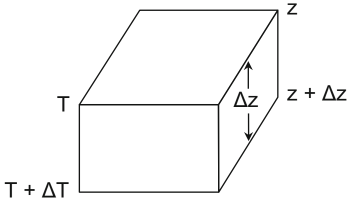


Fig. 6.8 Defining a temperature gradient

the upper mantle, with the iron possibly entering the lower mantle through the *D″ layer* that separates the lower mantle from the core.

Each model for mantle convection has strong supporters, and the question is not settled.

We consider next the variation of temperature as a function of depth within the Earth: $T = T(z)$.

6.1.2.4 Temperature Gradient

Figure 6.8 shows a horizontal slab of rock whose top surface is at a depth z in the Earth and whose bottom surface is at a depth $z + \Delta z$.

Here we adopt the convention that the depth, z, is zero at the Earth's surface and increases downward. That is, both z and Δz are positive downward.

The temperature of the top surface is T and that of the bottom surface is $T + \Delta T$. We expect temperature to increase inward in the Earth, so ΔT will be positive when Δz is positive. The ratio

$$\frac{\Delta T}{\Delta z}$$

gives the rate of change of temperature with depth at any fixed instant of time, and is called the *temperature gradient*. Because ΔT and Δz are both positive downward, it follows that the temperature gradient is positive downward.

In the limit as $\Delta z \to 0$, we have

$$\lim_{\Delta z \to 0} \frac{\Delta T}{\Delta z} = \frac{\partial T}{\partial z} \tag{6.7}$$

The use of partial derivative notation on the right side of (6.7) means that we are taking the derivative of T with respect to z while holding time fixed; that is, we are comparing the temperatures at two different points at the same instant of time.

By way of a counterexample, suppose a piece of rock is moving downward, as in mantle convection. This situation is illustrated in Fig. 6.9.

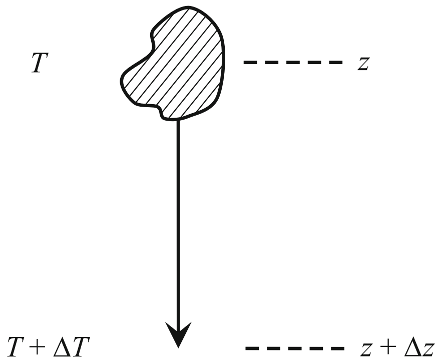


Fig. 6.9 Convective temperature gradient

Then the total derivative,

$$\frac{\mathrm{d}T}{\mathrm{d}z}$$

gives the rate of change of T with z for this rock as it moves; that is, the temperature is being compared at two different points at two *different* instants of time (the instants when the rock occupies those two points). So the full derivative does not depend on the depth alone, but also on time; while the partial derivative with respect to the depth depends only on the depth. In the language of mathematics, Fig. 6.9 demonstrates that $\mathrm{d}T/\mathrm{d}z$ depends on z and t.

6.1.3 Heat Conduction

Energy transport in the Earth's crust is by conduction.

Define the heat flux, Q, as the amount of heat energy passing through each square meter of surface area per second. The units of heat flux are therefore J m^{-2} s^{-1} or W/m^2.

Q is defined to be positive in the direction of increasing z, viz., downward, in the convention we have adopted here (see Fig. 6.10).

It turns out that, for a fixed separation Δz, $|Q|$ is larger for a larger temperature difference, ΔT; but, for a given ΔT, $|Q|$ is smaller if the surfaces are further apart (Δz is larger):

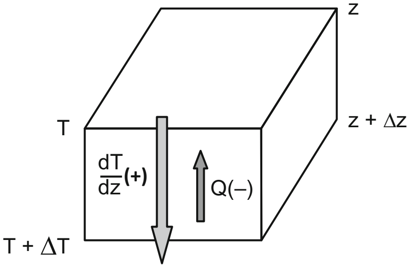
$$|Q| \propto \Delta T, \quad \textit{but inversely} \quad \propto \Delta z$$

Therefore, $|Q| \propto \frac{\Delta T}{\Delta z}$; and, in the limit as $\Delta z \to 0$,

$$|Q| \propto \frac{\partial T}{\partial z}$$

The constant of proportionality, k, is defined such that:

Fig. 6.10 Sign convention for Q and dT/dz



$$Q = -k\frac{\partial T}{\partial z} \tag{6.8}$$

where k is the *coefficient of thermal conduction,* and the minus sign is there because heat always flows from the hotter region toward the cooler region; that is, in the direction opposite to the temperature gradient (as shown in Fig. 6.10). The units of k are given by $Q/(\Delta T/\Delta z)$, i.e., (W/m^2)/(K/m) = Wm^{-1} K^{-1}.

Equation (6.8) is the one-dimensional form of a more general expression known as *Fourier's Law*. Equation (6.8) will, however, suffice for our treatment here. This is our *thermal conductivity equation.* A typical value of Q for crustal rocks on the Earth's surface is ~3 Wm^{-2}.

6.1.4 Energy Generation

Heat energy is generated inside the volume of rock because the radioactive decay of uranium, thorium, etc., releases heat. There may also be other sources or sinks of heat, such as pressure changes, latent heat released by phase changes, etc.

Define A = heat energy generated per cubic meter per second (W/m^3).

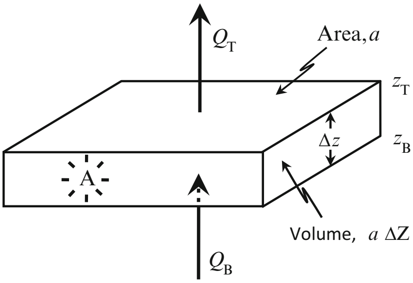
6.1.5 Equilibrium

The Earth and the other terrestrial planets have had time to lose most of their heat of formation (primordial heat), so we expect rocks in the crust to be in thermal equilibrium, on average. *Thermal equilibrium* means that rate of energy loss = rate of internal energy generation.

This is actually a quasi-equilibrium condition because, for example, the radioactive decay rate is decreasing very slowly with time; but these changes are slow enough that they can be ignored compared to the timescale of heat conduction.

Each individual volume inside the crust should also be in thermal equilibrium (on average), so the net energy gained by the volume over any length of time, δt, should be zero:

Fig. 6.11 The energy budget in *a* slab



The heat energy through bottom surface (positive if into the volume)
plus
The heat energy through top surface (negative if out of the volume)
plus
the heat generated in the volume
= 0

Figure 6.11 illustrates the situation.

The individual terms for the heat flux are, for the heat energy up through the bottom:

$$\begin{aligned}\text{heat energy through bottom} &= \begin{bmatrix}\text{heat energy per unit area} \\ \text{per unit time} \times \text{area} \times \text{time}\end{bmatrix} \\ &= -Q_B\, a\ \ \delta t\end{aligned} \tag{6.9}$$

The minus sign comes from the circumstance that the volume gains energy (LHS positive) with an upward heat flow (Q_B negative). The net result is that the quantity ($-Q_B$) is positive.

And, for the heat energy through the top,

$$\begin{aligned}\text{heat energy through top} &= \begin{bmatrix}\text{heat energy per unit area} \\ \text{per unit time} \times \text{area} \times \text{time}\end{bmatrix} \\ &= -Q_T\, a\, \delta t\end{aligned} \tag{6.10}$$

The volume loses energy (LHS negative) with an upward heat flow (Q_T negative), so the sign in the equation is positive.

Now for the heat generated within the volume:

$$\begin{aligned}\text{Heat generated in } \Delta V &= \begin{bmatrix}\text{heat generated per unit volume} \\ \text{per unit time} \times \text{volume} \times \text{time}\end{bmatrix} \\ &= A\, \Delta V\, \delta t = A\, a\, \Delta z\, \delta t\end{aligned} \tag{6.11}$$

These three terms have to add up to zero net energy gain, so,

$$-Q_{\mathrm{B}}\,a\ \ \delta t + Q_{\mathrm{T}}\,a\ \ \delta t + A\,a\,\Delta z\,\delta t = 0 \tag{6.12}$$

Dividing through by a and δt and rearranging terms gives

$$Q_{\mathrm{B}} - Q_{\mathrm{T}} = A\,\Delta z \tag{6.13}$$

or

$$\frac{Q_{\mathrm{B}} - Q_{\mathrm{T}}}{\Delta z} = \frac{Q_{\mathrm{B}} - Q_{\mathrm{T}}}{z_{\mathrm{B}} - z_{\mathrm{T}}} = A \tag{6.14}$$

Therefore, in the limit as $\Delta z \to 0$,

$$\frac{\partial Q}{\partial z} = A \tag{6.15}$$

If we use the thermal conduction equation (6.8) in (6.15) and divide through by –k, then we arrive at

$$\frac{\partial^2 T}{\partial z^2} = -\frac{A}{k} \tag{6.16}$$

Equation (6.16) can be integrated to find T as a function of z in the crust, where heat transport is by conduction, provided A and k are known at depth z.

However, A and k are not obtainable from seismic work in the manner that we can find v_{P} and v_{S}, the P- and S-wave velocities, discussed in Chap. 5. Thus, the geotherms are much more uncertain than the run of density with depth.

6.1.6 Central Temperature of the Earth

If energy were transported by conduction throughout the Earth, integration of (6.16) would be valid and would yield a temperature for the center of the Earth of about 60,000 K. This extremely high temperature results from the fact that conduction is a slow process, so a large temperature gradient is needed to produce the observed heat flow. The faster the temperature increases with depth into the Earth, the higher the central temperature will be.

Convection is a much more efficient process, because the warmer rocks are transported to cooler regions (carrying their heat with them) in a much shorter time than it would take the same heat to diffuse to the new region by conduction. The observed heat flow can then be produced by a much smaller temperature gradient, giving a central temperature for the Earth near 6,000–8,000 K.

6.2 Geotherms

A *geotherm* is an equation (or a table, or a curve on a graph) which gives temperature as a function of depth into the Earth. This can be created readily if we make certain simplifying assumptions. One of these involves adiabatic conditions. An *adiabatic process* is any process in which heat energy does not enter or leave the material involved. The temperature may change due to compression or expansion, but under adiabatic conditions, there is no exchange of heat energy *between* the material and its surroundings.

The speed of convection in the Earth is high compared to the rate of diffusion of heat by thermal conduction, so heat does not have time to enter or leave the moving material and the convection is adiabatic.

The temperature gradient in material which is convecting adiabatically is

$$\frac{\partial T}{\partial r} = -\frac{T\alpha g}{c_P} \tag{6.17}$$

where α is the coefficient of thermal expansion; c_P the specific heat at constant pressure; g the acceleration due to gravity; and T the temperature.

Integration of (6.16) for conduction gives the temperature and rate of heat flow at the base of the crust (top of the mantle). Equation (6.17) can then be integrated numerically from the top of the mantle to the center of the Earth (the outer core is also convecting), giving a geotherm for the whole Earth.

Typical results are shown in Fig. 6.12. Approximate uncertainties in the mantle temperatures in Fig. 6.12 are indicated by the width of the geotherm; and in the core by the dotted lines, based on the uncertainties in the temperatures at the inner and outer boundaries of the outer core. The temperature gradient is very steep at the core-mantle boundary, and is left as unknown in the figure.

The uncertainty results from the uncertainties in α and c_P at each depth. The data are based on laboratory experiments which measure the properties of iron and various rock mixtures at high temperatures and pressures. From this process, the temperature at the center of the Earth is determined to be 6,900 K ± 1,000 K.

6.3 Solar Heating

From (4.6) and (4.7), the luminosity of the Sun, $\mathcal{L}_\odot$, in terms of its radius, $R_\odot$, and its effective temperature, $T_\odot$, is

$$\mathcal{L}_\odot = 4\pi R_\odot^2 \sigma T_\odot^4 \text{ (W)} \tag{6.18}$$

and from Sect. 4.4.4, it is also expressible in terms of the solar constant and total sphere area at the Earth's orbit, (4.13):

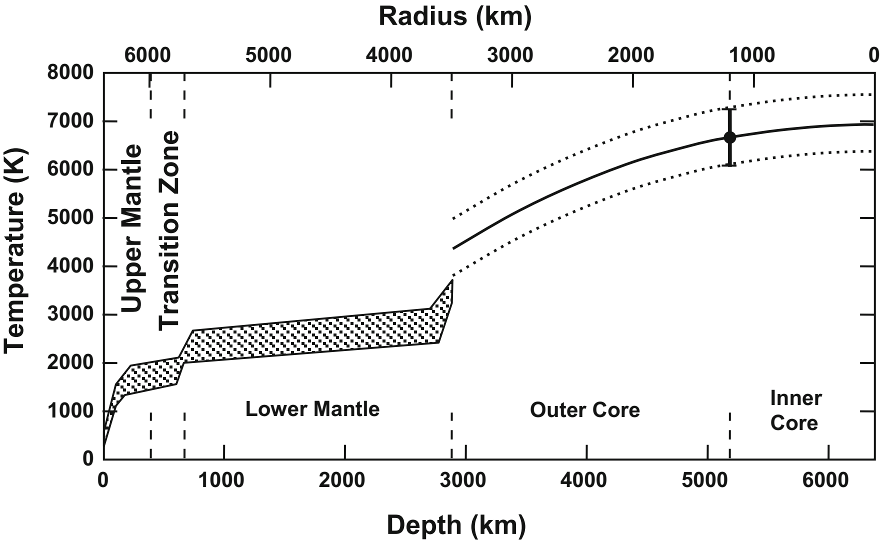


Fig. 6.12 Geotherm for the Earth. Mantle temperatures are from Jeanloz and Morris (1986) and references therein. The geotherm in the core is an approximate interpolation between three points: 4,400 K ± 600 K at the core-mantle boundary (Jeanloz and Morris 1986); 6,670 K ± 600 K at the inner core-outer core boundary (Alfè et al. 1999); and $\leq$6,900 K at the Earth's center (Williams et al., 1987), taken here as 6,900 K. The first two of these temperatures are consistent with those of Williams et al. (1987): 4,800 K ± 200 K and 6,600 K, respectively, and agree within uncertainty with the more recent determination by Anzellini et al. (2013) of 4,050 K ± 500 K and 6,230 K ± 500 K, respectively. We thank Prof. Bukowinski (1999; private communication 2013) for references to earlier material

$$\begin{aligned}\mathcal{L}_{\odot} &= 4\pi r^2 \mathfrak{S} = 3.826(8) \times 10^{33}\,\text{erg/s}\\ &= 3.826(8) \times 10^{26}\,\text{W}\end{aligned}$$

(Sect. 4.4.3; Allen 1973, p. 169). The flux at distance r from the Sun's center is then

$$\mathscr{F}(r) = \mathcal{L}_{\odot}/\left[4\pi r^2\right] \quad \left(\text{W/m}^2\right) \tag{6.19}$$

The power, P, striking any cross-sectional area, α, normal to the direction of radiation (e.g., an area of a planetary surface with the Sun at the zenith) is

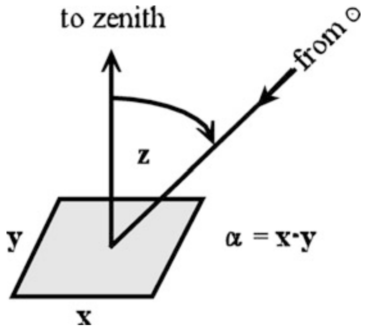
$$P = \alpha\mathscr{F} \quad (\text{W}) \tag{6.20}$$

and the power absorbed over that area is

$$P_{\alpha} = P(1 - A) = \alpha\mathscr{F} \bullet (1 - A) \quad (\text{W}) \tag{6.21}$$

where A is the *bolometric albedo* (effectively the ratio of reflected to incident bolometric flux). This power must be reradiated, because otherwise, the energy would increase continuously with time and the temperature would rise without limit

Fig. 6.13 Geometry of insolation



(contrary to both observation and the laws of thermodynamics!). The reradiated or emitted power for that area is analogous to the luminosity of a star:

$$P_e = \alpha\sigma T^4 \quad (\mathrm{W}) \tag{6.22}$$

where T is the equilibrium temperature. Setting $P_a = P_e$, so that

$$\alpha\mathscr{F}(1-A) = \alpha\sigma T^4 \tag{6.23}$$

$$T^4 = \frac{\mathscr{F}(1-A)}{\sigma} \tag{6.24}$$

and we get the equilibrium temperature:

$$T = \left[\frac{\mathscr{F}(1-A)}{\sigma}\right]^{1/4} = \left[\frac{\mathrm{L}_\odot(1-A)}{4\pi\sigma r^2}\right]^{1/4} \tag{6.25}$$

Note that the cross-sectional area cancels out. However, the equilibrium temperature at the subsolar point (i.e., the spot on the Earth where the Sun is overhead) is much higher than the equilibrium temperature of the planet as a whole, and is higher than the equilibrium temperature at sites where the zenith distance of the Sun exceeds 0°.

Figure 6.13 demonstrates the geometry when the Sun impinges on the area through a zenith distance, z, in which case (6.21) becomes:

$$\begin{aligned} P_\mathrm{a} &= P(1-A) \\ &= \alpha \cos z\, \mathscr{F}(1-A) \quad (\mathrm{W}) \end{aligned} \tag{6.26}$$

The factor α cos z arises because, if the area is xy, and the Sun, the x-axis, and the zenith lie in the same plane, the input flux is spread over a cross-sectional area:

$$(x \cos z)y = \alpha \cos z$$

The emission, however, is from the total area, $\alpha = xy$.

Again, setting the absorbed and emitted power, $P_a = P_e$, equal,

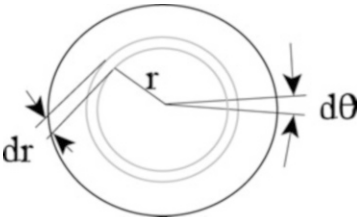


Fig. 6.14 Integration over the planetary disk, centered on the subsolar point

$$\alpha \cos z \ \mathscr{F} \ (1 - A) = \alpha\sigma T^4 \tag{6.27}$$

and designating the equilibrium temperature of (6.25) as $T_{\alpha,0}$ and that of (6.27) as $T_{\alpha,z}$, we get:

$$T_{\alpha,z} = T_{\alpha,0}[\cos z]^{1/4} \tag{6.28}$$

Looked at in this way, we see that $T_{\alpha,0}$, the temperature at the subsolar point (the point where the Sun is directly overhead), is the maximum possible temperature achievable during the day as the planet rotates. At any other place on the planet the maximum temperature is found at $z = z_{\min} = h_{\max}$, the minimum zenith distance and maximum altitude of the Sun at local noon, when the Sun is on the observer's *celestial meridian* (see Sect. 2.1.4).

These temperatures are purely *local* temperatures, in the sense that every part of the planet can be thought of as having a different local temperature. If the Sun is not in the sky at all, the local temperature will decrease as the reradiated energy matches the reduced heat flux into the designated area.

We may make a similar calculation for the planet as a whole.

From the standpoint of the radiation impinging on the planet, the planet appears as a flat disk. Designating an incrementally small area, d*a*, of this disk of radius $\mathfrak{R}$, we obtain for the absorbed power,

$$P_a = \int \mathscr{F}(1 - A) \ \mathrm{d}a \ \ (\mathrm{W}) \tag{6.29}$$

where the integration is taken over all sunlit areas. The integral can be taken over disk segments of width dθ and thickness dr, where r is the apparent distance of the area element from the subsolar point. Thus,

$$a = \int d\theta \int r \, dr \tag{6.30}$$

where the integration in theta is taken over 2π and the radius is taken from 0 to $\mathfrak{R}$, the radius of the planet. The integration over ring segments of radius r, width dr, and angular wedge dθ within the planetary disk is illustrated in Fig. 6.14.

Because $\int \mathrm{d}\theta = 2\pi$ and $\displaystyle\int r\mathrm{d}r = \frac{1}{2}\Re^2$, and with the assumption that we can assign a mean or effective bolometric albedo to the planet as a whole, we arrive at the equation for the area of a circle of radius $\Re$ times the solar flux at the planet and the average fraction of absorption, $<1 - A>$:

$$P_a = \pi\Re^2\mathscr{F} < 1 - A > \text{ (W)} \tag{6.31}$$

The reradiated power depends on how well the flux is distributed over the planet, i.e., it depends on the rotation as well as the atmospheric convection of the solar heat energy. Ignoring the latter, we can discuss two cases: *rapid* and *slow* rotations. An example of the latter would be a planet that is locked into its orbital angular rate so that the rotation period is equal to the revolutionary period (there is no such planet in the solar system). Such a planet would reradiate from half its surface area, whence the emitted power becomes:

$$P_e = 2\pi\Re^2\sigma T_s^4 \text{ (W)} \tag{6.32}$$

so that

$$T_\mathrm{s} = \left[\frac{\mathscr{F} < 1 - A >}{2\sigma}\right]^{1/4} = \left[\frac{\mathcal{L}_\odot < 1 - A >}{8\pi\sigma r^2}\right]^{1/4} \tag{6.33}$$

This is then an estimate of the mean equilibrium temperature of the sunlit hemisphere of a slowly rotating planet, or at any rate a planet rotating so slowly that a relatively insignificant amount of thermal emission is coming from the night side of the planet. The detection of some thermal radiation from the night side of Mercury in the 1960s proved conclusively that Mercury was not locked in a 1:1 spin–orbit coupling with the Sun; Doppler radar mapping later showed that the lock-in rate is, rather,

$$3P_{rotation} = 2P_{revolution}.$$

However, there could well be extra-solar planets, such as the "hot Jupiters," that fully satisfy the "slow rotator" conditions because of their extremely close-in orbits (see Milone and Wilson 2014, Chap. 16).

If the planet is a *rapid* rotator, on the other hand, the emitted power, from (6.22), becomes:

$$P_e = 4\pi\Re^2\sigma T_r^4 \text{ (W)} \tag{6.34}$$

because all of the planet now contributes to the emission, so that, on average,

$$T_\mathrm{r} = \left[\frac{\mathscr{F} < 1 - A >}{4\sigma}\right]^{1/4} = \left[\frac{\mathcal{L}_\odot < 1 - A >}{16\pi\sigma r^2}\right]^{1/4} \tag{6.35}$$

Note that in neither (6.33) nor (6.35) is there a dependence on the radius of the planet.

The relationship between these mean equilibrium temperatures is

$$T_{\mathrm{s}} = T_{\mathrm{r}} \cdot 2^{1/4} \tag{6.36}$$

In (6.22), we assume that the planet radiates as a black body (See Sect. 4.4.1), even though in (6.21) it does not absorb as a black body (because not all incident radiation is absorbed). This is not unreasonable for a planet orbiting a solar-type star, because the radiation incident from the star peaks at visible wavelengths whereas the radiation emitted by the much cooler planet peaks in the infrared. The albedo is different in the two wavelength regions. T, T_{s} and T_{r} in (6.25), (6.33) and (6.35) are then *effective temperatures*; i.e., the temperatures that the planet would have if it radiated as a black body.

If the planet does not radiate as a black body, then we can find the "true" temperature by replacing $\sigma T_{\mathrm{eff}}^4$ in (6.22), (6.27), (6.32) and (6.34) with $\varepsilon\sigma T_{\mathrm{true}}^4$, where ε is called the *emissivity*[2].

If ε were known, the RHS of (6.33) and (6.35) could be divided by $\varepsilon^{1/4}$ to find the temperature in terms of mean albedo and emissivity, the distance from the Sun, and, of course, the radiation properties of the Sun.

Example 6.1 Here we compute the Earth's equilibrium temperature.
$\mathcal{L}_{\odot} = 3.83 \times 10^{26}$ W, $\mathscr{F} = 1{,}362$ Wm^{-2}, $A_{\oplus} = 0.307$, and, with $\Re_{\oplus} = 6.378 \times 10^6$ m, the absorbed power is, from (6.31),

$$P_a = \pi\Re^2 \mathscr{F} < 1 - A >= \ 1.21 \times 10^{17}\mathrm{W}.$$

Because the Earth can be assumed to be a rapidly rotating planet, we have, for the emitted power, from (6.34),

$P_e = 4\pi\Re^2\sigma T^4$, and, therefore, the effective equilibrium temperature for this (rapid rotation) case, is

$$T_{\mathrm{r}} = \left[\frac{\mathcal{L}_{\odot} < 1 - A >}{16\pi\sigma r^2}\right]^{1/4} = \left[\frac{3.83 \times 10^{26} \times 0.693}{16\pi \times 5.67 \times 10^{-8} \cdot \left[1.496 \times 10^{11}\right]^2}\right]^{1/4}$$
$$= 254.0 \ \mathrm{K}$$

For more examples, see Schlosser et al. (1991/1994, Chap. 17, 18). Lewis (1995/1997, Chap. 5) has a calculation for the giant planets.

Usually, equilibrium temperatures differ from actual observed temperatures for many reasons, such as clouds(!), thermal inertia on the sunset side of a slowly rotating planet, atmospheric circulation effects, internal heat sources and the greenhouse effect. To account for "thermal inertia", or intermediate rotation cases, the factor 2π in (6.32) and 4π in (6.34) can be replaced by 2π f, where $1 \le \mathrm{f} \le 2$ is a reradiation factor. A planetary atmosphere acts to scatter radiation from the Sun (and other astronomical objects!), which contributes to the local heating budget.

[2] N.B: This emissivity is not that defined in (4.22).

Also, it is well known that H_2O and CO_2 are important *greenhouse* gases, which strongly absorb in the IR preventing complete cooling by radiation from the surface and lower atmosphere (an extremely important condition on Venus but also present on the Earth). Moreover, atmospheric molecules cool the (upper) atmosphere through the emission process by radiating into space. A source of cooling in the Earth's atmosphere at 20 km altitude, for example, is the radiation due to the 15 μm line of CO_2. These processes are dealt with in detail in Goody (1995).

The presence of oceans on Earth is thought to have played a major role in the regulation of CO_2 and the development of photosynthetic vegetation, which consumes CO_2 and produces O_2 and H_2O. The carbon dioxide-consuming plants have played a further critical role in the evolution of Earth's atmosphere into a nitrogen–oxygen one. In Milone and Wilson, (2014, Chaps. 10 and 12) we discuss the structure of the atmosphere, circulation effects, and the properties of individual planets.

This concludes our general discussion of the heating and temperature of a terrestrial planet. We return next to the rocks and minerals of these planets' interiors, concentrating once more on the material of that planet we know best, the Earth.

Challenges

[6.1] At what solar zenith angle is the local temperature equal to the global equilibrium temperature of a rapidly rotating planet (assuming black body radiation is solely responsible for the temperatures).

[6.2] Suppose internal heat were the only source of heat flux to the Earth's surface. What would be the equilibrium temperature of the Earth?

[6.3] Suppose the heat flux from the interior were to equal the average heat flux from the Sun, and that it was uniform around the Earth. What equilibrium temperature would result? Would this calculation need to be refined further because of the effects of a higher temperature on the Earth's surface and atmosphere?

[6.4] Suppose the rotation equator of a planet were inclined 90° to its orbital plane. Would such a planet be characterized as a slow rotator or a rapid one? Assume a circular orbit and compute the equilibrium temperature at two critical points in its orbit to demonstrate your finding.

References

Allen, C.W. (ed.): Astrophysical Quantities, 3rd edn. The Athlone Press, University of London, London (1973)

Anderson, D.L.: Theory of the Earth. Blackwell Publications, Boston (1989)

Anzellini, S., Dewaele, A., Mezouar, M., Loubeyre, P., Morard, G.: Melting of iron at earth's inner core boundary based on fast X-ray diffraction. Science **340**, 464–466 (2013)

Alfè, D., Gillan, M.J., Price, G.D.: The melting curve of iron at the pressures of the Earth's core from ab initio calculations. Nature **401**, 462–464 (1999)

Arevalo Jr., R., McDonough, W.F., Luong, M.: The K/U ratio of the silicate Earth: insights into mantle composition, structure and thermal evolution. Earth Planet. Sci. Lett. **278**, 361–369 (2009)

Bukowinski, M.S.T.: Earth science: taking the core temperature. Nature **401**, 432–433 (1999)

Davies, J.H., Davies, D.R.: Earth's surface heat flux. Solid Earth **1**, 5–24 (2010)

Davis, E.E., Chapman, D.S., Mottl, M.J., Bentkowski, W.J., Dadey, K., Forster, C., Harris, R., Nagihara, S., Rohr, K., Wheat, G., Whiticar, M.: FlankFlux: an experiment to study the nature of hydrothermal circulation in young oceanic crust. Can. J. Earth Sci. **29**, 925–952 (1992)

Gando et al. (The KamLAND Collaboration, 66 authors): Partial radiogenic heat model for Earth revealed by geoneutrino measurements. Nat. Geosci. **4**, 647–651 (2011)

Goody, R.: Principles of Atmospheric Physics and Chemistry. University Press, Oxford (1995)

Jeanloz, R., Morris, S.: Temperature distribution in the crust and mantle. Annu. Rev. Earth Planet. Sci., **14**, 377–415 (1986)

Lay, T., Hernlund, J., Buffett, B.A.: Core–mantle boundary heat flow. Nat. Geosci. **1**, 25–32 (2008)

Lewis, J.S.: Physics and Chemistry of the Solar System. Academic, San Diego, CA (1995/1997)

Milone, E.F., Wilson, W.J.F.: Solar System Astrophysics: Planetary Atmospheres and the Outer Solar System. Springer, New York (2013)

Schlosser, W., Schmidt-Kaler, T., Milone, E.F.: Challenges of Astronomy. Springer, New York (1991/1994)

Schubert, G., Turcotte, D.I., Olson, P.: Mantle Convection in the Earth and Planets. Cambridge University Press, Cambridge (2001)

Stacey, F.D., Davis, P.M.: Physics of the Earth, 4th edn. Cambridge University Press, Cambridge (2008)

Stein, C., Lowman, J.P., Hansen, U.: The influence of mantle internal heating on lithospheric mobility: implications for super-earths. Earth Planet. Sci. Lett. **361**, 448–459 (2013)

Stephenson, F.R., Morrison, L.V.: Long-term fluctuations in the earth's rotation: 700 B.C. to A.D. 1990. Trans. R. Phil. Soc. Lond. **A351**, 165–202 (1995)

Williams, Q., Jeanloz, R., Bass, J., Svendsen, B., Ahrens, T. J.: "The Melting Curve of Iron to 250 Gigapascals: A Constraint on the Temperature at Earth's Center." Science, **236**, 181–182 (1987)

Wood, B.: The formation and differentiation of Earth. Phys. Today **64**, 40–45 (2011)

Chapter 7
Rocks and Minerals

The fundamental components of terrestrial planets, rocky moons, asteroids and meteoroids are rocks and minerals. Therefore, an understanding of rocks and minerals is essential to understanding rocky objects in any planetary system. In this chapter we review basic mineral and rock types that are common, to one extent or another, to all such objects, and how these minerals are altered by pressure and temperature and thus with depth into a planet. Starting in Sect. 7.6, we apply these concepts to the interior of the Earth in a manner that allows comparison with other planets in Chap. 9. We begin by making some basic distinctions concerning rocks and minerals.

7.1 Rocks

Rocks differ from minerals in that minerals are crystalline solids having a definite chemical composition (e.g., quartz (SiO_2)), and rocks are consolidated assemblages of minerals. Rocks can be classified as igneous, sedimentary, or metamorphic, as follows.

7.1.1 Igneous Rocks

Igneous rocks are those that solidified directly from molten rock, in the form of either magma or lava. *Magma* is molten rock below the planetary surface, whereas *lava* is magma that comes out onto the surface.

The sizes of crystal grains in igneous rocks depend on the rate of cooling. If the rock cools quickly then there is no time for large crystals to form, and the crystals are small or absent. If the rock cools slowly then the crystals are larger.

E.F. Milone and W.J.F. Wilson, *Solar System Astrophysics: Background Science and the Inner Solar System*, Astronomy and Astrophysics Library,
DOI 10.1007/978-1-4614-8848-4_7, © Springer Science+Business Media New York 2014

Igneous rocks are divided into three types, based on grain size:

1. *Extrusive, or volcanic, rocks* formed from lava. Being on the surface, they cooled quickly and therefore have a very fine grain or no grain structure. They are often porous (containing pores or pockets formed by the release of dissolved gases) and are often associated with volcanic out throws or lava flows.
2. *Intrusive rocks* formed from magma at shallow depths below the surface. Because they remained covered, they cooled more slowly, and have small to medium grain sizes.
3. *Plutonic rocks* formed from magma which remained deep below the surface. They cooled very slowly, over millions of years, and have a coarse grain structure and low porosity.

7.1.2 Sedimentary Rocks

Sedimentary rocks are formed after igneous and other rocks are eroded (weathered), carried away by wind or water, and deposited as sediments. If these sediments become overlain by other sediments, they can eventually be compressed to form rock.

Some examples of sedimentary rocks on the Earth are:

1. *Limestone*: Formed from calcium carbonate ($CaCO_3$) from plankton that died and settled onto the ocean floor. Limestone is the main constituent of the Rocky Mountains.
2. *Sandstone*: Formed from quartz (SiO_2) or feldspar sand grains that were transported by wind or water and became cemented together under pressure by materials such as silica, clays and calcite. Quartz and feldspar are discussed later in this chapter.
3. *Mudstone*: Formed from mud, composed primarily of clay minerals, that has been compacted into rock. If the mudstone is in thin layers (laminated) that can be split apart (fissile), it is referred to as *shale*.

7.1.3 Metamorphic Rocks

Metamorphic rocks are formed when sedimentary or igneous rocks are metamorphosed (changed) by high temperature and pressure, without melting.

Some examples of metamorphic rocks on the Earth are:

1. *Marble*: Formed from compression and heating of limestone or other carbonate rock. Recrystallization destroys the original, sedimentary carbonate structure to give a uniform structure of interlocking crystals.
2. *Quartzite*: Formed from compression and heating of quartz sandstone. The quartz grains, together with the cementing material, are recrystallized to form a dense, hard rock that fractures through the quartz grains, whereas sandstone fractures around the quartz grains.

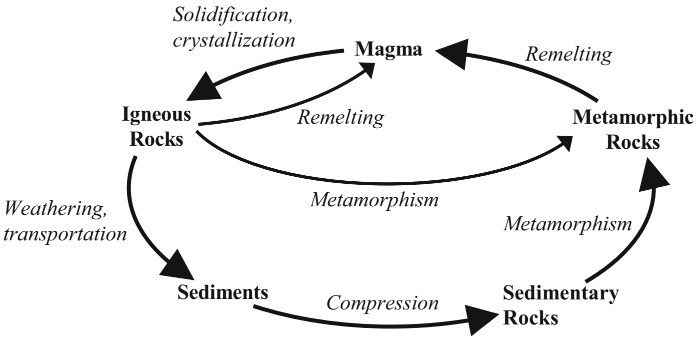


Fig. 7.1 The geochemical cycle

3. *Schist and gneiss*: Formed from compression of sedimentary rocks such as shales, or from igneous rocks. The layering is produced by shearing forces that align mica crystals and other platy minerals in planes perpendicular to the direction of compression.

The average composition of the Earth's crust is approximately 94 % igneous rock, 6 % sedimentary rock, and <1 % metamorphic rock.

7.1.4 The Geochemical Cycle

The *geochemical cycle* is shown schematically in Fig. 7.1. Heavy lines show the main cycle, in which magma solidifies to form igneous rocks that are subsequently weathered and compressed to form sedimentary rocks. These undergo metamorphism to produce metamorphic rocks, and the metamorphic rocks are remelted to form magma. Many other paths are also possible, two examples of which are shown by lighter lines.

7.2 Minerals

Each different mineral has its own chemical composition (e.g., $MgSiO_3$) and/or crystal structure (e.g., face-centered cubic) and/or crystal size (e.g., large, small, absent), etc.

7.2.1 Crystal Structure

There are many possible structures for crystals. We will describe two of these in detail here.

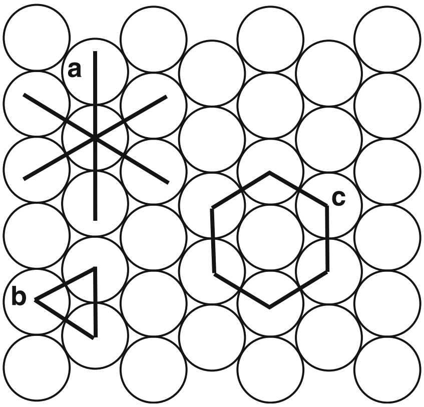


Fig. 7.2 A single layer of closely packed spheres

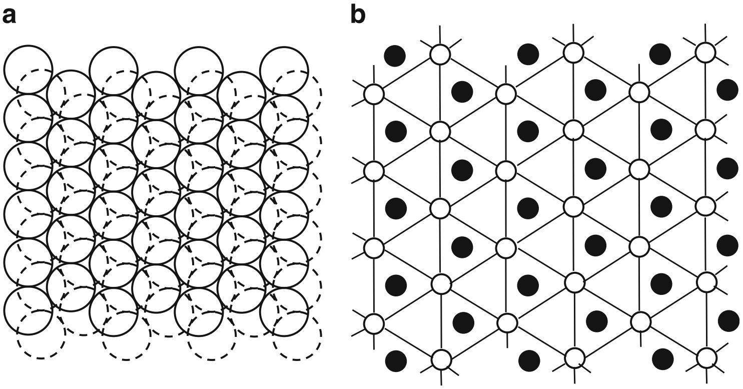


Fig. 7.3 Two layers of closely packed spheres with the upper layer resting as low as possible over the lower layer (see text)

We begin by packing spheres of equal size into a single layer in the densest possible manner, as shown in Fig. 7.2. Three properties of this close-packing arrangement are:

1. The spheres form a three-way grid (lines 120° apart), as indicated by the lines labeled "a" in Fig. 7.2.
2. Any three spheres in mutual contact form an equilateral triangle (e.g., the triangle labeled "b" in Fig. 7.2).
3. Every sphere has six equidistant neighbors touching it, forming a hexagon ("c" in Fig. 7.2).

If a second, identical, layer is stacked on top of this one, then the central planes of the two layers will be closest together when the second layer sits as low as it can over the first layer. This happens when the spheres in the upper layer lie over the holes (centers of the equilateral triangles) in the lower layer. Figure 7.3a shows the resulting arrangement. Here, solid circles represent the upper layer and dashed

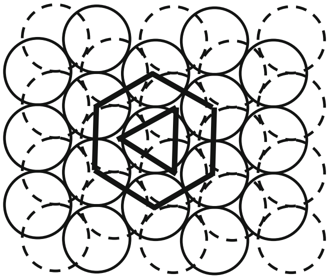


Fig. 7.4 A small section of Fig. 7.3a

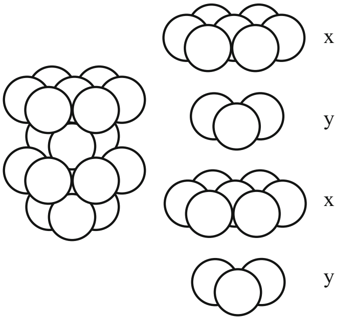


Fig. 7.5 Perspective (*left*) and exploded (*right*) views of four consecutive layers in hexagonal close packing (HCP)

circles represent the lower layer. Figure 7.3b shows a more open view (with a slightly expanded scale) which may make the arrangement clearer. Here, only the centers of the spheres are shown. Open circles indicate spheres in the upper layer, and filled circles indicate spheres in the lower layer.

A careful examination of Fig. 7.3 shows that some of the holes in the upper layer lie over *spheres* in the lower layer and others lie over *holes* in the lower layer. The fact that there are two types of holes in the upper layer means that there are now two different ways to stack a third (identical) layer on top of this one.

1. Stack the third layer so that its spheres are directly over the *spheres* in the layer two below it. If we continue stacking in this fashion, then the layers in Fig. 7.3 will alternate in an xyxyxy. . . pattern. The solid circles on the left side of Fig. 7.3 then represent ions in layers 1, 3, 5. . . and the dashed circles represent ions in layers 2, 4, 6....

   Figure 7.4 shows a portion of Fig. 7.3. Consider the hexagon of solid circles in Fig. 7.4 alternating with the triangle of dashed circles. The triangles are, of course, also parts of hexagons. Figure 7.5 shows (on the left) a three-dimensional view of four successive layers of this arrangement and (on the right) an exploded view of the same four layers.

   This method of stacking layers of spheres is called *hexagonal close packing* (HCP).
2. Stack the third layer so its spheres are over *holes* in *both* of the two layers below it. If we continue stacking in this fashion, then the spheres in any given layer are over the spheres in the layer *three* below, and the pattern of stacking that results is xyzxyzxyz....

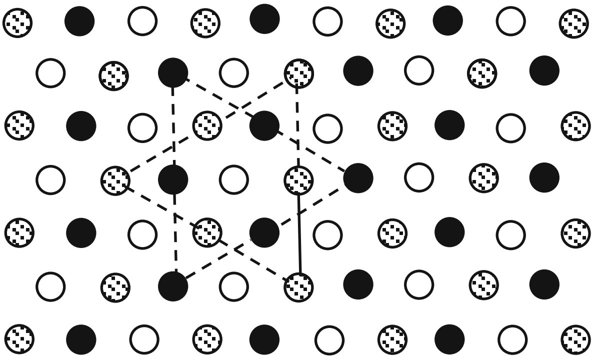


Fig. 7.6 A view through a crystal in which the layers are stacked xyzxyzxyz…, as explained in the text

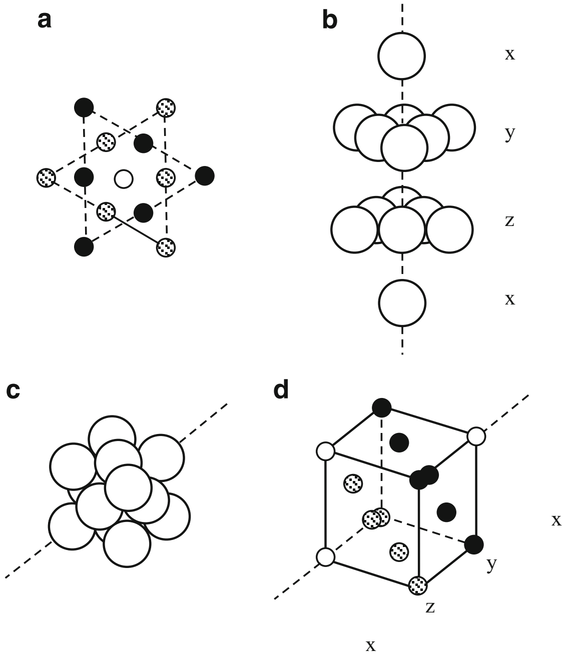


Fig. 7.7 Arrangement of layers and ions in cubic close packing (CCP), producing a face-centered cubic (FCC) crystal

Figure 7.6 shows a view through such a crystal (compare with the right-hand diagram in Fig. 7.3).

As we look down through the crystal, open circles represent layers 1, 4, 7,…, filled circles represent layers 2, 5, 8,…, and stippled circles represent layers 3, 6, 9,… The star-shaped arrangement shown in Fig. 7.6 is seen again in Fig. 7.7.

Other arrangements are also possible, such as xyzyxyzyx…, but these do not concern us here.

Figure 7.7a shows the star-shaped portion of Fig. 7.6. Figure 7.7b shows an exploded three-dimensional view of four successive layers in this portion. There are a total of 14 ions in these four layers. Figure 7.7c shows these same four layers pushed together as they exist in the crystal; and Fig. 7.7d shows schematically the locations of the 14 ions involved. Figure 7.7c, d have been rotated to the right and away from the viewer compared to Fig. 7.7b, as indicated by the dashed line (which is the same line in all three diagrams).

Figure 7.7c, d show that the 14 ions make up a cube. Eight ions form the corners of the cube, and the other six ions are located at the centers of the six faces. This xyzxyz . . . pattern is therefore referred to as *cubic close packing* (CCP), and forms a *face-centered cubic* (FCC) lattice.

Note that the layering formed by the top and bottom faces and the face-centered ions in the FCC pattern (Fig. 7.7c, d) is not the same as the layering we used to generate the pattern (Fig. 7.7a, b). The relationship between the two is shown by the arrangement of open, filled, and stippled circles in Fig. 7.7d, which correspond to the same symbols in the upper left-hand diagram, and in Fig. 7.6.

7.2.2 Crystal Density

If we pack steel balls in a box, then the balls in each layer are not directly affected by the balls in the layer two below. As a result, the layers are the same distance apart whichever arrangement (xyxyxy or xyzxyz) is used, and the overall density (g/cm^3 or kg/m^3) is the same for both HCP and CCP (or FCC).

However, ions of the same sign (+ or –) repel each other even if they are not physically touching. Therefore, an ion in a given layer is repelled more by the layer two below if the ion in the upper layer is over an ion in the lower layer (xyxyxy) than if it is over a hole in the lower layer (xyzxyz). As a result, a CCP (FCC) crystal has a greater density in kg/m^3 than an HCP crystal made of the same ions.

7.2.3 Interstitial Holes

When one layer is stacked on top of another to form a pair of layers in contact, holes (spaces) are created between the ions in one layer and the ions in the other layer. They are not the same as the "two-dimensional" holes discussed above, which are at the centers of the equilateral triangles *in* each layer. If we think of the midplanes of the layers as being located at $x = 0$ and $x = 1$, then for the close-packed layers discussed here the holes will be near $x = 1/2$. These holes between two layers are referred to as *interstitial holes*.

Interstitial holes have a three-dimensional structure, in that each hole is surrounded by a small, three-dimensional framework of ions. There are two types of interstitial holes, depending on the number of ions (shape of the framework) surrounding the hole:

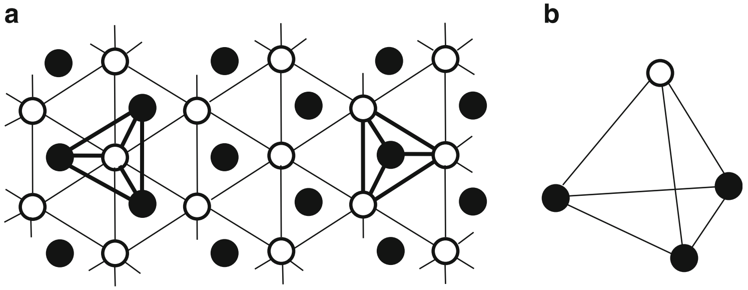
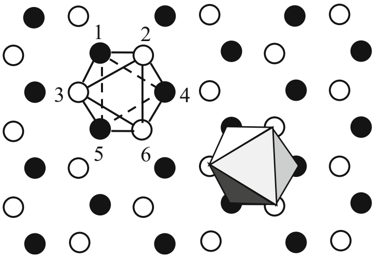


Fig. 7.8 Tetrahedral holes between adjacent layers of a crystal. *Open circles* represent ions in the upper layer and *filled circles* represent ions in the lower layer

Fig. 7.9 Octahedral holes between adjacent layers of a crystal. *Open circles* represent ions in the upper layer and *filled circles* represent ions in the lower layer



7.2.3.1 Tetrahedral Holes

If an equilateral triangle of ions in the lower layer has an ion above it, then the hole between the two layers is surrounded by four ions, all in mutual contact. The hole is then at the center of a *tetrahedron* : a four-sided structure, each face of which is an equilateral triangle.

Two examples of tetrahedral holes are shown by heavy lines in Fig. 7.8a, with a three-dimensional view in Fig. 7.8b. (Remember that open and filled circles represent only the centers of the ions; the ions themselves are larger and in mutual contact.) The left-hand tetrahedron in Fig. 7.8a points out of the page, and the right-hand one points into the page.

7.2.3.2 Octahedral Holes

If the equilateral triangle in the lower layer does *not* have an ion over it in the layer above, then the hole between the layers is surrounded by *six* ions.

Two examples are shown in Fig. 7.9. In the left-hand example, the ions surrounding the hole are labeled from 1 to 6. The other example has been shaded to bring out the three-dimensional structure of the ions surrounding the hole more clearly.

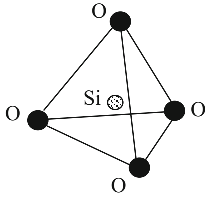


Fig. 7.10 Silicate tetrahedron. The silicon ion (*stippled circle*) occupies the tetrahedral hole at the center of the tetrahedral framework of four oxygen ions (*filled circles*)

Two of the ions in the lower layer (1 and 5) form a square with two in the upper layer (2 and 6). One other ion in the upper layer (3) and one in the lower layer (4) lie "above" and "below" the midpoint of the square, along its central axis.

As a result, the hole is at the center of an octahedron, an eight-sided figure made of two four-sided pyramids base-to-base.

7.2.4 The Silicate Tetrahedron

The crystals of silicate minerals (described below) can be regarded as stacked layers of oxygen ions, with silicon ions occupying a fraction of the tetrahedral sites and other metal ions occupying a fraction of the octahedral sites. The respective fractions of tetrahedral and octahedral sites occupied are different for different minerals.

The silicate tetrahedron (SiO_4) is shown in Fig. 7.10, and is the basic building block of almost all silicate minerals. (In the case of stishovite, perovskite and the post-perovskite phase, discussed below, the basic building block is a silicate octahedron.)

The valence of silicon is +4 and that of oxygen is –2, so the valence of SiO_4 is $4 + 4(-2) = -4$. Each SiO_4 therefore bonds easily with other ions, usually those of silicon or metals.

7.2.5 Mineral Names

Mineral names are somewhat arbitrary, because:

1. Minerals with the same crystal structure but different chemical compositions can have the same name. This happens because certain types of ions (e.g., magnesium and iron) are so similar to each other in size and chemical properties that one can replace the other in the octahedral sites with no change to the crystal structure.

   An example is olivine, a silicate mineral in which each occupied octahedral site contains an ion of either magnesium (Mg) or iron (Fe). If all such sites are occupied by Mg, then the mineral is called *forsterite* (Mg_2SiO_4), and if all are occupied by iron then the mineral is called *fayalite* (Fe_2SiO_4).

   However, these are simply end-members of a continuous series of possible compositions, and the chemical symbol for olivine is often given as

Table 7.1 Mean composition of the Earth's crust

Element	Symbol	Weight %	Atom %
Oxygen	O	46.60	60.5
Silicon	Si	27.72	20.5
Aluminum	Al	8.13	6.2
Iron	Fe	5.00	1.9
Calcium	Ca	3.63	1.9
Sodium	Na	2.83	2.5
Potassium	K	2.59	1.8
Magnesium	Mg	2.09	1.4
Titanium	Ti	0.44	
Hydrogen	H	0.14	
Phosphorus	P	0.12	
		99.29	

$(Mg_{(1-x)}Fex)_2$ SiO_4, where x can have any value from 0 to 1. For example, $(Mg_{0\ 9}Fe_{0.1})_2SiO_4$ has 90 % of the sites occupied by magnesium and 10 % by iron. All these possible compositions are included in the name olivine.

Olivine is described more fully in Sect. 7.2.8.4.

2. Minerals with the *same* chemical composition can exist in different *phases* (different crystal lattices) and these can have different names.

   Olivine and silicate spinel (hereafter referred to simply as "spinel") provide an example. Both have the chemical symbol Mg_2SiO_4 (or Fe_2SiO_4, etc), but they differ in crystal structure: in olivine the oxygen ions are arranged in an HCP lattice and in spinel they are arranged in a CCP (or FCC) lattice.

   Spinel is a high-pressure phase of olivine. Olivine is common in the Earth's upper mantle, but deeper in the mantle where pressures are higher it undergoes a *phase transition* to the more closely packed crystal structure of spinel. Spinel therefore has the same chemical symbol as olivine, but is about 10 % denser. Spinel is discussed again in Sects. 7.2.8.4 and 7.4.2.

   A more familiar example of a substance changing name when it undergoes a phase transition is, of course, water changing to ice.

3. Some minerals have different names even though they have the same chemical composition *and* crystal structure, because they differ in crystal size or some other characteristic.

7.2.6 Composition of the Earth's Crust

The mean chemical composition of the Earth's crust, excluding some trace constituents, is given in Table 7.1.

That is, oxygen atoms make up 60.5 % of all atoms in the crust, but, because oxygen atoms are lighter than all of the others listed except hydrogen, they make up only 46.60 % of the mass or weight of the crust.

Oxygen and silicon together make up 74 % of the weight of the crust, or 81 % of the atoms. It is therefore clear that the most common minerals will be those involving oxygen and silicon: *the silicates.*

7.2.7 Oxygen-to-Silicon Ratio in Chemical Symbols

The basic crystal unit of most silicate minerals is the SiO_4 tetrahedron. If the tetrahedra are isolated from each other (i.e., they are not linked by sharing oxygen atoms), then the mineral will contain four oxygen atoms for every silicon atom and the chemical symbol will contain SiO_4, e.g., olivine, Mg_2SiO_4.

However, if some oxygen atoms are shared between adjacent tetrahedra (i.e., the tetrahedra are linked), then on average there are fewer oxygen atoms per silicon atom in the mineral. This sharing changes the chemical symbol, even though the basic crystal unit is still the SiO_4 tetrahedron; e.g., pyroxene, $MgSiO_3$; quartz, SiO_2.

7.2.8 Oxygen-to-Metal Ratio and the Classification of Silicate Minerals

One way to classify silicate minerals is by their oxygen-to-metal ratio.

The most common metals in silicate minerals are silicon, magnesium, iron, calcium, potassium, and sodium, but others are also possible.

The most common ratios are, in order of decreasing oxygen abundance (or increasing metal abundance):

$$\text{oxygen} : \text{metal} = 2{:}1 \quad 8{:}5 \quad 3{:}2 \ \text{and}\ 4{:}3$$

7.2.8.1 2:1 Ratio: Silica (SiO_2)

Silica has different mineral names, depending on the crystal structure. Two examples are quartz and stishovite.

Quartz is a framework silicate: not all oxygen ions in the close-packing arrangements discussed above are actually present, and the remaining tetrahedra form a framework.

Figure 7.11 shows a single layer of linked tetrahedra in a quartz crystal. Large circles represent oxygen ions, and small circles represent silicon ions (dashed if hidden below an oxygen ion). The bases of the tetrahedra share oxygen ions to form a repeating pattern of hexagons, while the tetrahedra themselves point alternately up and down around each hexagon.

The upward-pointing tetrahedra in this layer share their oxygen ions with the downward-pointing tetrahedra in the next layer.

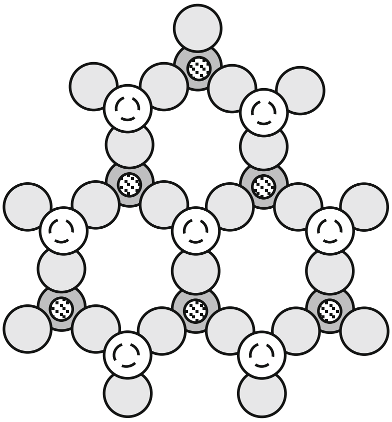


Fig. 7.11 A single layer of linked tetrahedra in a quartz crystal

The SiO_4 tetrahedra therefore link to form a three-dimensional framework in which each tetrahedron shares all four oxygen ions with its neighbors. The result is that, on average, there are two oxygen ions for every silicon ion, and the chemical symbol for silica is SiO_2, even though it is made up of linked SiO_4 tetrahedra (or linked SiO_6 octahedra).

Stishovite is a high-pressure (~20 GPa, where 1 GPa = 1 gigapascal = 10^9 Pa = 10 kbar) phase of quartz, made up of SiO_6 octahedra. The oxygen ions are shared between adjacent octahedra so that on average there are again two oxygen ions for every silicon ion, and the chemical symbol is still SiO_2.

7.2.8.2 8:5 Ratio: Feldspar Group

$M_2AlSi_2O_8$, where M_2 = two metal ions. One of the metal atoms in the chemical symbol is potassium, K, sodium, Na, or calcium, Ca, and the other is aluminum, Al, or silicon, Si. Three examples of feldspars are:

$$\begin{aligned} &\text{Orthoclase feldspar}: && KAlSi_3O_8 \\ &\text{Plagioclase feldspar}: && \begin{cases} NaAlSi_3O_8 \text{ (albite)} \\ CaAl_2Si_2O_8 \text{ (anorthite)} \end{cases} \end{aligned}$$

Feldspars are framework silicates like quartz. For example, in orthoclase the Al^{3+} replaces one of every four Si^{4+} in the tetrahedral sites in quartz, and the charge imbalance is compensated by introducing K^+ ions into octahedral sites.

7.2.8.3 3:2 Ratio: Pyroxene Group

$M_2Si_2O_6$, where M_2 = two metal ions. The metal atoms are commonly Ca, Mg, or Fe, or less commonly K, Al, Ti, or Na.

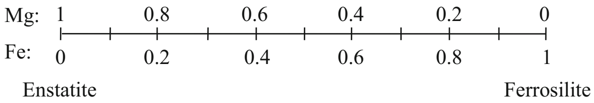


Fig. 7.12 Enstatite–ferrosilite series

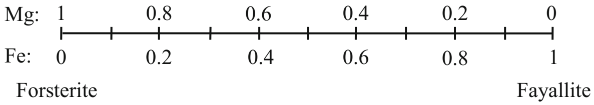


Fig. 7.13 Forsterite-fayalite series

Two examples of pyroxenes are enstatite, $MgSiO_3$ (equivalent to $Mg_2Si_2O_6$) and diopside, $CaMgSi_2O_6$.

As illustrated in Fig. 7.12, enstatite forms one end-member of a continuous series of possible compositions with ferrosilite, $FeSiO_3$, in which iron replaces the magnesium of the enstatite. The enstatite–ferrosilite series is referred to as (Mg,Fe) SiO_3. For example, the pyroxenes in the upper mantle have a composition near $(Mg_{0.9}Fe_{0.1})SiO_3$, which means that 90 % of the available spaces are filled by magnesium ions and 10 % by iron ions.

The $MgSiO_3$–$FeSiO_3$ series is referred to as *orthopyroxene,* and the $CaMgSi_2O_6$–$CaFeSi_2O_6$ series as *clinopyroxene.*

The pyroxenes are chain silicates: each tetrahedron shares a basal oxygen ion with the next to form a zigzagging chain stretching along the *c* direction. Because of the shared O ions, counting along a chain produces a sequence of one Si, three O, one Si, three O, etc.; so the silicate portion of the chemical symbol is SiO_3. Parallel chains form planes parallel to the *b,c* plane, and adjacent planes are separated (and bonded together) in the *a*-direction by the positive metal cations. All tetrahedra in a given chain have their apices pointing in the same direction, either upward or downward relative to the *b,c* plane. If the apices of one chain point "up," then those of the adjacent chains in both the *a* and *b* directions point down, with chains alternating in this fashion through the crystal.

Phase transitions in pyroxene at high pressures can produce garnet, perovskite, and the post-perovskite phase (i.e., a phase at a higher pressure than perovskite), described in Sects. 7.4.1.2, 7.4.1.4, and 7.4.1.5, respectively. These also have an oxygen:metal ratio of 3:2, and also have a range of compositions.

7.2.8.4 4:3 Ratio: Olivine and Spinel

M_2SiO_4, where M_2 = two metal ions (magnesium or iron). *Olivine* has a continuous range of compositions, referred to as $(Mg, Fe)_2SiO_4$, in which forsterite, Mg_2SiO_4, and fayalite, Fe_2SiO_4, are the two end-members. The series is illustrated in Fig. 7.13.

In olivine, the oxygen ions form an HCP lattice in which the SiO_4 tetrahedra are isolated from each other (that is, they do not share oxygen ions). The metal ions

(Mg or Fe) occupy octahedral sites between the tetrahedra and bond the silicate tetrahedra together in the *x, y* and *z* directions.

Spinel is a high-pressure phase of olivine, formed at pressures >12 GPa.

Olivine and spinel are described again in Sect. 7.4.2.

7.3 Mineral Content of Igneous Rocks

Table 7.2 lists the mineral content of igneous rocks.

In Table 7.2 we note the following:

1. The *amphibole* group consists of chain-type silicate minerals in which every second SiO_4 tetrahedron in a chain shares an oxygen ion with a tetrahedron in an adjacent chain, forming a double-chain structure. Adjacent double chains in amphibole are bonded together by positively charged ions, as is the case with the single chains of the pyroxenes.
2. *Pegmatite* is an uncommon rock that contains uranium, thorium, and other elements with unusual ion radii. These are the last to enter minerals in a solidifying magma, and therefore become concentrated in the crust.
3. *Eclogite* is obtainable by phase transitions from basalt (e.g., through subduction of basaltic crust). Eclogite may be one of the rock types making up the mantle.
4. *Pyrolite* (not listed in Table 7.2) is defined as any substance which can produce basaltic magma through partial melting, leaving behind dunite and peridotite as a result. Pyrolite in this picture is regarded as a single source material for both the basalt in the crust and the dunite and peridotite in the upper mantle, and is therefore itself an important constituent of the upper mantle.

   Pyrolite is a "fictitious" rock, in that it is defined in terms of what it does, leaving its actual mineral content to be determined. It seems to be composed of 2/3 olivine and 1/3 pyroxene, plus perhaps garnet.
5. *Serpentinite* (not listed in Table 7.2) is composed mainly of olivine and pyroxene combined with water. The water lowers the melting point significantly. Serpentinite is mentioned here only to show that the presence of water complicates the description of melting or crystallization of rocks in the Earth's upper mantle.

Table 7.2 Mineral content of igneous rocks

Density (g/cm)	Plutonic rock	Volcanic rock	Principal mineral content
	Pegmatite		Quartz, feldspar
2.7 - - - - - - - - - -	Granite	Rhyolite	Quartz, feldspar
	Diorite	Andesite	Feldspar, amphibole
3.0 - - - - - - - - - -	Gabbro	Diabase, basalt	Feldspar, pyroxene
	Peridotite		Pyroxene, olivine
3.5 - - - - - - - - - -	Dunite		Olivine
	Eclogite		Pyroxene, garnet

Table 7.3 Felsic/mafic classification

Plutonic rock	Volcanic rock	Felsic/mafic	Acidity
Granite	Rhyolite	Felsic	Acidic
Diorite	Andesite	Intermediate	
Gabbro	Diabase, basalt	Mafic	Basic
Dunite, eclogite		Ultramafic	Ultrabasic

7.3.1 Classification of Rocks by Mineral Content

Rocks may be classified according to their feldspar, silica, magnesium, and iron content.

Granite and rhyolite are silica-rich, as shown by the fact that quartz is one of the principal minerals in each one (see Table 7.2). They are also rich in feldspar, so they have a relatively high content of potassium and sodium. Such rocks are classified as *felsic* because of their high *fel*dspar and *s*ilica content. They are also more *acidic* than basalt or dunite.

Gabbro and basalt contain less silica and more magnesium, iron, and calcium than felsic rocks, and are classified as *mafic* rocks, from the words *ma*gnesium and *f*err*ic*. They are *basic,* rather than acidic.

Rocks such as dunite and eclogite, which have an even higher magnesium and iron content and are even more basic, are often referred to as *ultramafic* or *ultrabasic.*

These classifications are shown in Table 7.3.

7.4 Phase Transitions

A phase of a mineral is a particular crystal structure for that mineral.

As pressure increases with depth into the mantle, minerals undergo phase transitions in which the atoms (ions) are rearranged into a different crystal structure; for example:

$$\mathrm{Mg_2SiO_4\ olivine} \quad \rightarrow \quad \mathrm{Mg_2SiO_4\ spinel},$$

where the HCP lattice of oxygen ions in the olivine phase is transformed to the denser CCP lattice in the spinel phase.

This rearrangement can also involve one crystal structure separating into two crystal structures (i.e., the original mineral separates into two minerals) or two crystal structures merging into one (two minerals merging into a single mineral); for example,

$$\mathrm{Mg_2SiO_4\ spinel} \quad \rightarrow \quad \mathrm{MgSiO_3\ perovskite} + \mathrm{MgO\ periclase}$$

MgO has the same crystal structure as NaCl (table salt), as described in Sect. 7.4.1.4.

7.4.1 *Phase Transitions of Pyroxene, $MgSiO_3$*

7.4.1.1 Pyroxene

This is the low-pressure phase, found on a planet's surface and in the crust, e.g., enstatite, $MgSiO_3$.

The crystal structure of enstatite consists of chains of corner-sharing silicate tetrahedra. Because of the sharing of oxygen ions there are, on average, three oxygen ions (valence –2) for each silicon ion (valence +4), so the valence of each SiO_3 unit (one silicon ion plus the base of a tetrahedron) is $3 \times (-2) + 4 = -2$. The valence of Mg and Fe is +2, so on average there is one Mg or Fe ion for each SiO_3 unit, giving the chemical symbol $(Mg,Fe)SiO_3$.

As described in Sect. 7.2.8.3, pyroxene has a variety of compositions. For example, $MgSiO_3$ enstatite forms a continuous series with $FeSiO_3$ ferrosilite by progressive replacement of Mg ions with Fe ions; or every second Mg ion in $MgSiO_3$ enstatite can be replaced by a calcium (Ca) ion, forming $CaMgSi_2O_6$ diopside.

7.4.1.2 Garnet

Pyroxene undergoes a phase transition to silicate garnet, $Mg_4Si_4O_{12}$, at about 17–19 GPa pressure and $T > 2{,}000$ K, as shown in Fig. 7.17.

During the phase transition, the silicate tetrahedra change from corner-sharing chains (pyroxene, silicate unit = SiO_3) to isolated tetrahedra (garnet, silicate unit = SiO_4). In the process, one silicon ion is "released" from a tetrahedron, producing five metal ions for every three silicate tetrahedra in garnet:

$$4\left(\mathrm{SiO_3} + \mathrm{Mg}\right) \rightarrow 3(\mathrm{SiO_4}) + \mathrm{Si} + 4\mathrm{Mg}$$

That is, in $Mg_4Si_4O_{12}$ three of the silicon ions are inside SiO_4 tetrahedra ($3\ SiO_4 = Si_3O_{12}$), and the other silicon and the four magnesium make up the five metal ions occupying spaces between the tetrahedra.

There is a similar range of compositions for garnet as for pyroxene. Also, pyroxene in the Earth's crust and mantle is often mixed with Al_2O_3 (corundum). One Mg^{2+} ion in the garnet can then be replaced by Al^{3+} from the corundum. This produces a charge imbalance which can only be compensated if one Si^{4+} ion is also replaced by Al^{3+}, giving aluminous garnet, $Mg_3Al_2Si_3O_{12}$ (pyrope):

$$\begin{array}{l} \mathrm{Mg_4Si_4O_{12} + Al_2O_3 \ \rightarrow\ Mg_3Al_2Si_3O_{12}\ +\ MgSiO_3} \\ (\text{silicate garnet} + \text{corundum} \rightarrow \text{pyrope} + \text{pyroxene}) \end{array}$$

At the pressures and temperatures found in the upper mantle, aluminous garnet (pyrope) is also soluble in silicate garnet, producing aluminous silicate garnet (majorite, first synthesized in the lab by Ringwood and Major (1971)).

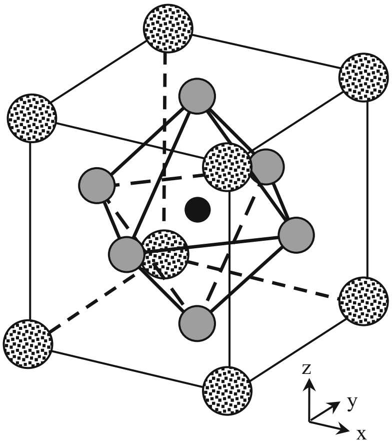


Fig. 7.14 Basic structural unit of a perovskite crystal. *Stippled circles*: Mg; *shaded circles*: O; *black circle*: Si

The crystal structure of corundum consists of alternating layers of oxygen and aluminum. The oxygen forms an HCP lattice, and the aluminum occupies distorted octahedral sites between the layers of oxygen.

7.4.1.3 Silicate Ilmenite

Above 20 GPa, $Mg_4Si_4O_{12}$ garnet undergoes a phase transition to a form of $MgSiO_3$ with the same crystal structure as $FeTiO_3$ ilmenite. This form of $MgSiO_3$ is generally referred to as silicate ilmenite. The crystal structure is the same as corundum (Al_2O_3), described above, with the Mg^{2+} and Si^{4+} replacing the Al^{3+} in an ordered alternate fashion.

Silicate ilmenite exhibits the same continuous sequence of compositions from $MgSiO_3$ to $FeSiO_3$ as do pyroxene and garnet.

7.4.1.4 Perovskite

Above 23 GPa, $MgSiO_3$ ilmenite undergoes a phase transition to a perovskite structure, $MgSiO_3$.

Figure 7.14 shows a basic structural unit of a perovskite crystal. The Mg^{2+} (stippled) and O^{4-} (shaded) ions form a face-centered cubic lattice with the Mg at the corners and the O at the centers of the faces. The Si^{4+} ion (black) is located at the center of the cube. Figure 7.14 shows that the oxygen ions form an octahedron with the silicon at its center.

When we abut other cubes against this one, they all share adjoining faces and corners. Consequently, each O ion is shared between two cubes and each Mg ion is shared between eight cubes. The perovskite crystal thus consists of linked SiO_6 octahedra in which each octahedron shares all six corner oxygen ions with its neighboring octahedra, and the magnesium ions occupy the spaces between the SiO_6 octahedra.

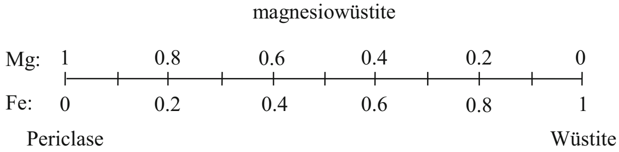


Fig. 7.15 Periclase–wüstite series

Now imagine an infinitely repeating sequence of such cubes in all directions to form a perovskite crystal. Some thought will show that every octahedron has a Mg ion to its lower right (as viewed in Fig. 7.14), and that this statement accounts for all the Mg ions; thus there is one Mg ion for every SiO_6 octahedron in the crystal. Also, if we imagine walking along each of the x, *y*, and *z* axes indicated in Fig. 7.14, then starting from any silicon ion, along each axis we have one Si, one O, one Si, one O, etc. There are three axes, so there are three O ions for each Si in the crystal. The chemical symbol for silicate perovskite is thus $MgSiO_3$, despite the fact that the basic building block is an SiO_6 octahedron.

In similar fashion to olivine and pyroxene, perovskite forms a continuous series in which some Mg ions are replaced by Fe, giving $(Mg_{(1-x)}Fe_x)SiO_3$. This form of perovskite is probably the most abundant mineral in the entire Earth; it may make up more than 80 % by volume of the lower mantle.

Other minerals also have phases with perovskite structure at high pressure, such as $CaSiO_3$ and $CaMgSi_2O_6$ (diopside).

However, pure $FeSiO_3$ has no perovskite phase, because at high pressure it decomposes into stishovite and wüstite:

$$FeSiO_3 \rightarrow SiO_2\,(\text{stishovite}) + FeO(\text{wüstite}).$$

FeO (wüstite) and MgO (periclase) are two end-members of another continuous series, (Mg,Fe)O (magnesiowüstite), as indicated in Fig. 7.15.

Magnesiowüstite has the same crystal structure as table salt (NaCl). The larger, negative oxygen ions form a face-centered cubic close-packed lattice, with the smaller, positive Mg or Fe ions occupying the spaces. The Mg/Fe ions also form a face-centered cubic lattice.

7.4.1.5 Post-perovskite Phase

Two different groups in 2004 announced the discovery that, at a pressure of ~120 GPa, perovskite undergoes a phase transition to a new crystal structure referred to as *post-perovskite* (Murakami et al. 2004, Oganov and Ono 2004). As shown in Fig. 7.16, the structure consists of planes of SiO_6 octahedra stacked along the *b* axis with Mg^{2+} ions between the layers. Within each plane, the octahedra share O ions along their edges to form chains parallel to the *a* axis, and the chains are connected by shared apical O ions along the *c* axis. Because all O ions are shared, the chemical symbol is $MgSiO_3$, as for perovskite (Sect. 7.4.1.4). The Mg^{2+}

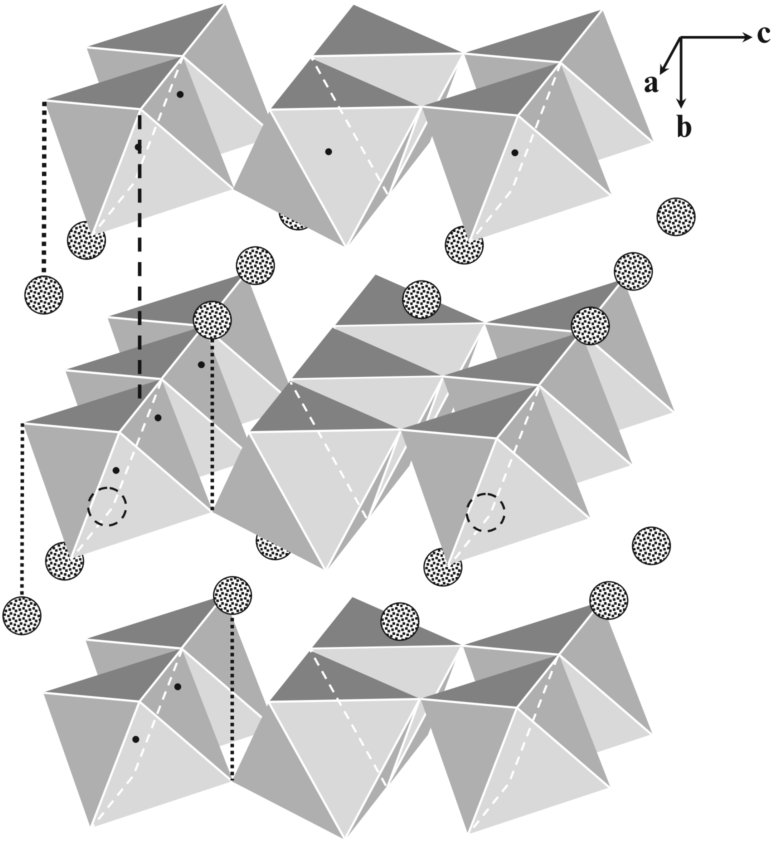


Fig. 7.16 Crystal structure of the post-perovskite phase of $MgSiO_3$ computed at 120 GPa pressure, derived from Figure 1 of Oganov & Ono (2004) and Figure 3 of Murakami et al. (2004). Stippled circles: Mg^{2+} ions (shown as a dashed circle if hidden behind an octahedron). Oxygen ions (not shown) are located at the corners of the octahedra, and each octahedron has a Si ion at its centre. The positions of some Si ions are indicated by black dots. The Mg^{2+} ions are located above or below apical O ions, as indicated by the four black dotted lines. Alternate planes of octahedra are displaced in the a-direction by half the width of an octahedron, as indicated by the black dashed line. Lines of Mg^{2+} ions are displaced in the a-direction by half the width of an octahedron relative to adjacent lines in both the b- and c-directions. Image drawn by W. J. F. Wilson from a handmade model

sites in post-perovskite are smaller than in perovskite, resulting in a volume reduction and corresponding density increase of 1.0–1.5 % (Hirose and Lay 2008).

7.4.1.6 Phase Diagrams for Pyroxene

Figure 7.17 shows a pressure-temperature phase diagram for pure $MgSiO_3$ for pressures and temperatures found in the Earth's mantle. As an example of phase transitions, the dashed, horizontal line illustrates the phase transitions from pyroxene to garnet to ilmenite to perovskite that would occur if pressure were to increase at a constant temperature of 2,000 °C.

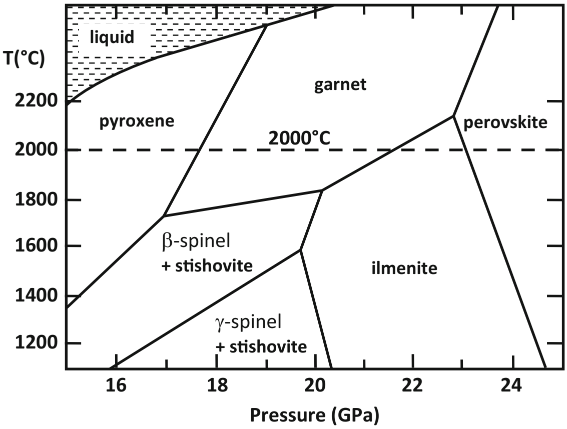


Fig. 7.17 Pressure-temperature phase diagram for $MgSiO_3$. Adapted from Anderson (1989), Fig. 16.4, p. 346, and appearing here with permission

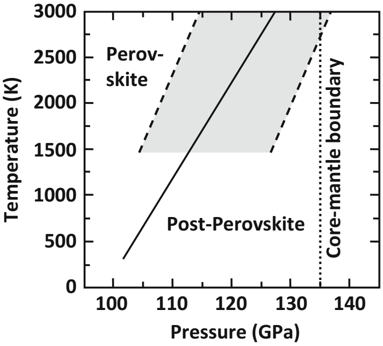


Fig. 7.18 Pressure-temperature phase diagram showing the computed perovskite-post-perovskite phase boundary for conditions in the lowermost ~800 km of the mantle. The position and slope of the phase boundary are somewhat uncertain, but the general appearance and strong, positive slope are as indicated. Two examples are shown, illustrating the range of computed slopes: *solid line*: Oganov and Ono (2004), Clapeyron slope 9.56 MPa/K; *dashed lines*: Catalli et al. (2009), Clapeyron slope 6.7 ± 0.5 MPa/K. Catalli et al. (2009) find a boundary width of ~20 GPa (*shaded area*) within which both phases are present

Figure 7.18 shows the perovskite-post-perovskite phase boundaries obtained by Oganov and Ono (2004) and Catalli et al. (2009) under lower-mantle conditions. The range in *Clapeyron slope* (the inverse of the slope in Fig. 7.18),

$$\frac{dP}{dT} = \frac{L}{T\Delta v} \tag{7.1}$$

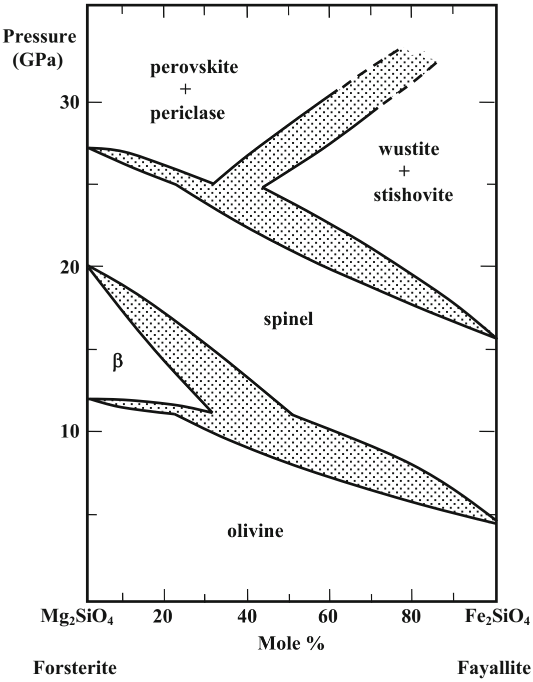


Fig. 7.19 Isothermal phase diagram at 1,000 °C for olivine (forsterite–fayalite system). From Bassett (1979), Fig. 6, p. 372, and reproduced here with permission

found for this transition is illustrated by the two boundaries shown: 9.56 MPa/K for Oganov and Ono (2004) and 6.7 ± 0.5 MPa/K for Catalli et al. (2009). In (7.1), L is the specific heat of transition (or latent heat), v is the specific volume (volume per unit mass, $v = 1/\rho$), and the equation assumes a sign convention in which heat input to the system is positive. If a sample of perovskite is compressed through the phase transition, then ρ increases and Δv is negative. Because the Clapeyron slope in Fig. 7.18 is positive, L in (7.1) must be negative; i.e., heat is released in the transition from perovskite to post-perovskite. The transition is therefore exothermic. In contrast, the ilmenite to perovskite transition (with a negative slope to the phase boundary in Fig. 7.17) is endothermic.

7.4.2 Phase Transitions of Olivine, Mg_2SiO_4

Figure 7.19 is an isothermal phase diagram of pressure versus iron fraction in the (Mg, Fe) part of olivine from 0 % (pure Mg_2SiO_4, forsterite) to 100 % (pure Fe_2SiO_4, fayalite), at a constant temperature of 1,000°C.

Olivine (also called α-phase) is the low-pressure phase. The crystal structure of olivine consists of isolated SiO_4 tetrahedra (not sharing oxygen atoms), with Mg or Fe in octahedral sites between the tetrahedra. The oxygen ions form an HCP lattice.

If the olivine is magnesium-rich, it undergoes a phase transition to β-phase (also called modified spinel) as pressure increases above about 20 GPa. Iron-rich olivine, however, has no β-phase and transforms directly into spinel.

β-Phase is denser than olivine, with a CCP (FCC) lattice of oxygen ions (see Sect. 7.2.1). The tetrahedra are in pairs that share an oxygen atom at one corner.

β-Phase (if magnesium-rich) or olivine (if iron-rich) transforms to a spinel structure at higher pressures. Spinel, also called γ-phase, has a CCP (FCC) lattice of oxygen ions with isolated tetrahedra, and is denser than either olivine or β-phase.

Many oxides crystallize with spinel structure, such $MgAl_2SiO_4$.

At higher pressures and temperatures, spinel decomposes into other minerals as shown in Fig. 7.19. It is interesting that, for the expected iron-to-magnesium ratio of about 90:10 in the mantle, both pyroxene and olivine produce perovskite at lower-mantle pressures.

7.5 Densities of Minerals

Table 7.4 lists the uncompressed densities (at 1 atmosphere pressure) of various minerals mentioned above.

7.6 Seismic Discontinuities in the Earth's Mantle

Figure 5.21 shows an empirical model of P- and S-wave speeds in the Earth's mantle, and the corresponding density profile, from Dziewonski and Anderson (1981). The three significant discontinuities in wave speed at depths of 220, 400, and 670 km below the Earth's surface all correspond to abrupt changes in density, and are believed to arise from phase changes in the solid rock with increasing depth (and therefore increasing pressure) in the mantle, as described below. The *Mohorovičić discontinuity*, or "*Moho*," marks the crust–mantle boundary and is believed to arise from a composition change from basaltic material in the crust to dunitic or peridotitic material in the upper mantle, rather than a phase change. Variations of depth or radius with latitude and longitude are not shown; e.g., the depth of the Moho varies from 5 km below oceanic crust to 35 km below continents and even 60 km or more below mountain ranges.

Table 7.4 Uncompressed densities of minerals (at 1 atm pressure) in kg/m^3

Olivine (Mg_2SiO_4) series		ρ	Pyroxene ($MgSiO_3$) series		ρ
Forsterite	Mg_2SiO_4	3,210	Enstatite	$MgSiO_3$	3,200
β-phase	Mg_2SiO_4	3,470	Garnet	$Mg_3Al_2Si_3O_{12}$	3,560
Spinel	Mg_2SiO_4	3,560	Ilmenite	$MgSiO_3$	3,800
Periclase	MgO	3,580	Perovskite	$MgSiO_3$	4,110
Stishovite	SiO_2	4,290			

7.7 Relationship of Phase Diagrams to Seismic Discontinuities

A model for the Earth's mantle is shown in Fig. 7.20 (Liu 1979), with the phase transitions described above plotted on a pressure scale (top horizontal axis) and also converted to depth into the Earth (bottom horizontal axis).

The density, ρ_o, plotted along the vertical axis in Fig. 7.20 is the uncompressed density, as in Table 7.4, not the actual density under pressure.

In this model, the 400 and 670 km seismic discontinuities are seen as arising from mineral phase transitions:

400 km discontinuity: olivine to β-phase

670 km discontinuity:

spinel to perovskite + periclase in the olivine component

and

garnet to ilmenite in the pyroxene/corundum component.

A small seismic discontinuity has sometimes been suspected between 400 and 650 km depth, and may result from the β-phase to spinel transition in the olivine component.

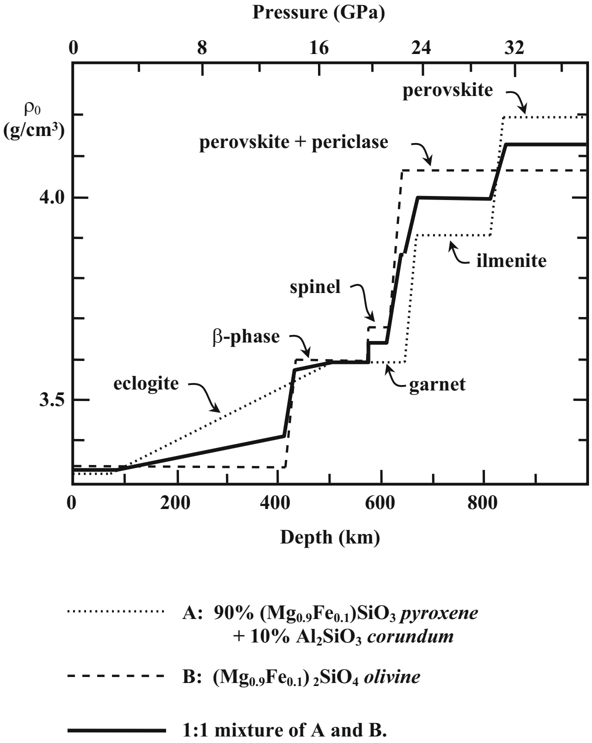


Fig. 7.20 Model of phase transitions in the mantle, with olivine, pyroxene, and corundum in the proportions given in the legend. From Liu (1979), Fig. 7, p. 195, with permission

Another small discontinuity below 670 km has also been suspected, and may be due to the transition from ilmenite to perovskite in the pyroxene/corundum component.

7.8 The Effect of Phase Boundaries on Mantle Convection

A phase boundary in a convecting region can have a significant effect on the convection, either intensifying it or inhibiting it.

Figure 7.21a illustrates two columns of rock in the mantle of a terrestrial planet, one convecting upward (left column) and the other downward (right column) through surrounding stationary rock. White indicates the less dense phase 1 and grey the more dense phase 2, with the phase boundary indicated by the dark line between them. As parcels of rock in the column rise or fall, the phase boundary within the column remains at a fixed depth (not necessarily the same depth as in the surrounding rock), and the rock changes phase as it passes through this boundary. Figure 7.21b shows a corresponding schematic phase diagram for the case where the transition from phase 1 to phase 2 is exothermic, so the slope of the phase boundary (solid line) is positive. The graph of temperature versus pressure for the surrounding rock (dashed line) crosses the phase boundary at point b. Pressure increases with increasing depth, so point b corresponds to a particular depth in the mantle (see Fig. 7.21a).

The effect of the phase boundary on the convection depends on the relative importance of three factors (Schubert et al. 1975): advection of temperature, latent heat, and thermal expansion. In the discussion below, we will consider these three factors separately, although in nature they occur together.

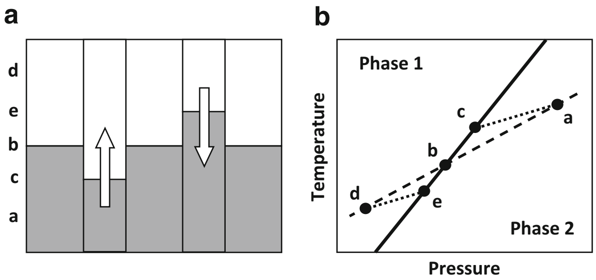


Fig. 7.21 Effect of phase boundaries on mantle convection. (**a**) Columns of rock convecting through surrounding stationary rock in the mantle in the presence of a phase boundary. *White*: Phase 1. *Grey*: Phase 2. The *arrows* show the direction of motion of each column. (**b**) Schematic phase diagram for the case of an exothermic phase change (heat is released) for a transition from the less dense phase 1 to the more dense phase 2. Greater pressure corresponds to greater depth. *Solid line*: Phase boundary. *Dashed line*: Temperature (T) versus pressure (P) in the surrounding rock. *Dotted line* (**a**–**c**): T versus P for a parcel of rock in the ascending column. *Dotted line* (**d**–**e**): T versus P for a parcel of rock in the descending column. Letters are explained in the text

1. Advection of temperature. Consider a parcel of rock in the rising column, located in phase 2 at some point, a, below the phase boundary. For purposes of discussion, we will assume that the parcel and the surrounding rock have the same temperature at point a. Heat takes time to diffuse between the parcel and its surroundings, so as the parcel rises into cooler rock it finds itself hotter than its surroundings. The parcel reaches the phase boundary (and changes phase) at point c. Because the phase boundary has a positive slope, point c is at a higher pressure (greater depth) than point b. The rock between points c and b in the rising column (see Fig. 7.21a) is then less dense than its surroundings and its buoyancy intensifies the convection. A parcel in phase 1, descending from point d, finds itself cooler than the surrounding rock and changes phase at point e, at a lower temperature and pressure and therefore smaller depth than in the surrounding rock. The descending rock between points e and b therefore finds itself denser than its surroundings, and its negative buoyancy also intensifies the convection. It is left as an exercise for the reader to show that, if the phase change from phase 1 to phase 2 is endothermic (negative slope to the phase boundary), then the displacements of the phase boundary are opposite to those in Fig. 7.21a, and advection of temperature inhibits convection.

2. Latent heat. If the transition is exothermic from phase 1 to phase 2, then it is endothermic from phase 2 to phase 1. The phase change at point c then cools the rising parcel, and it slides downward along the phase boundary in Fig. 7.21b as the transition proceeds. Conversely, the descending parcel is heated by the exothermic transition, and slides upward along the phase boundary. In both cases, the release or absorption of latent heat acts to oppose the effect of advection of temperature, inhibiting convection. It is left as an exercise for the reader to show that, if the transition from phase 1 to phase 2 is endothermic (negative slope), the release or absorption of latent heat adds to the effect of advection of temperature (which is to inhibit convection), thus further inhibiting convection. The release or absorption of latent heat therefore acts to inhibit convection for all cases.

3. Thermal expansion. If the transition from phase 1 to phase 2 is exothermic, then in the ascending (descending) parcel, the decrease (increase) in temperature resulting from the absorption (release) of latent heat causes the parcel to contract and become denser (expand and become less dense), inhibiting convection in both cases. If the transition from phase 1 to phase 2 is endothermic, then convection is intensified. Note that the effect in point 3 is opposite to that in point 1, above.

Thus, whether the transition from phase 1 to phase 2 is exothermic or endothermic, one factor amplifies convection and the other two inhibit it. Whether convection is amplified or inhibited in the combined effect can be determined only from quantitative stability computations; but as a rule exothermic phase transitions amplify convection and endothermic transitions inhibit it (Breuer et al. 1996).

7.9 The Core–Mantle Boundary and the D″ Layer

The D″ layer, shown schematically in Fig. 7.22, is a region about 200–300 km thick observed at the base of the lower mantle, in contact with the core. Its upper boundary is defined by a discontinuity in which in the S-wave velocity, v_S, increases by 2–3 % from the lower mantle to the D″ layer (Catalli et al. 2009; see also Hirose and Lay 2008 for a review of the D″ layer). Any discontinuity in the P-wave velocity, v_P, however, is generally much smaller or absent.

Globally, the depth of the S-wave discontinuity (and therefore the thickness of the D″ layer) varies laterally by 40–50 km, and the discontinuity may even disappear in places. Within the D″ layer, S-waves show an anisotropy in which the velocities, v_{SH}, of horizontally-polarized S-waves (SH waves) are ~1 % greater than v_{SV}, the velocities of vertically-polarized S-waves (SV waves) over most of the D″ layer (Panning and Romanowicz 2004). However, the anisotropy is reversed ($v_{SV} > v_{SH}$) at the base of two broad low-velocity regions, often called *superplumes*, in the mantle below the Pacific Ocean and Africa.

Seismic studies of the D″ layer indicate a heterogeneous composition, both radially and laterally, perhaps due to ancient material left over from mantle differentiation, subducted oceanic lithospheric slabs reaching the base of the mantle, and core-mantle chemical reactions, and also a possible narrow thermal boundary layer in contact with the core where heat flow is by conduction rather than convection (Hirose and Lay 2008). However, a more thorough understanding of

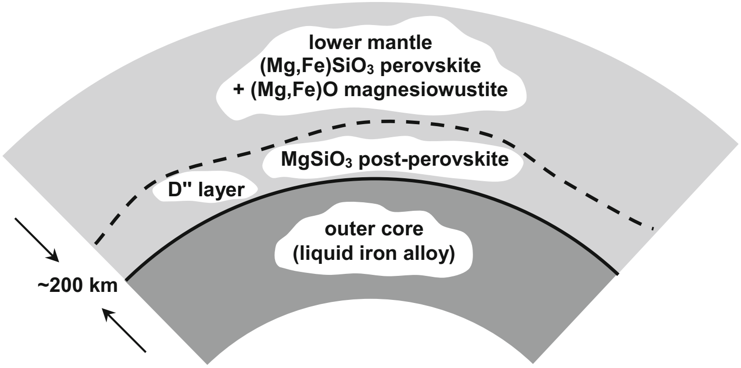


Fig. 7.22 The D″ layer, a layer of heterogeneous composition at the base of the upper mantle

many characteristics, particularly the S-wave anisotropy, had to await the discovery of the post-perovskite phase of pyroxene in 2004 (Sect. 7.4.1.5). Some points regarding this phase transition are (see Hirose and Lay 2008):

1. Computations and experiments at temperatures and pressures characteristic of the core-mantle boundary using laser-heated diamond-anvil cells show that the phase change creates a 1.4–4 % increase in v_S but only ±0.5 % change in v_P. Velocity changes of <0.5 % are difficult to measure, so both of these results are consistent with the observations mentioned above.

2. The large, positive Clapeyron slope in Fig. 7.18 means that the phase transition will occur at a shallower depth for cooler, subducting material and deeper for warmer, rising material in plumes (Fig. 7.21), contributing to the lateral depth variations described above.

3. The stacked plane structure of post-perovskite provides a natural explanation for the S-wave anisotropy described above. If there were large-scale lateral flow in some parts of the D″ layer and vertical flow in others, then the stacked planes would act as slip planes, and the crystals would tend to align with their slip planes parallel to the flow. They would also be more compressible along an axis perpendicular to the planes than within the planes. The S-wave velocity is given by

$$v_S = \sqrt{\frac{\mu}{\rho}} \tag{7.2}$$

   where μ is the rigidity of the material and ρ is the density, so S waves whose polarization axes lie in the crystal planes (where the rigidity is greater) would travel faster than those polarized perpendicular to the planes. The horizontal and vertical anisotropies then suggest that the D″ layer forms a boundary layer for mantle convection, with subducting material entering, plumes leaving, and lateral flow between.

4. Seismic observations often show an anticorrelation between S-wave velocity, v_S, and *bulk-sound-wave velocity*, v_C, in the D″ layer. The bulk wave velocity is given by

$$v_C = \sqrt{\frac{K_S}{\rho}} = \sqrt{v_P^2 - \frac{4}{3}v_S^2} \tag{7.3}$$

   where K_S is the bulk modulus. Calculations show that v_C is larger and v_S is smaller in post-perovskite than in perovskite, accounting at least in part for this anticorrelation.

In Fig. 7.18, the positive slope of the phase boundary means that the phase boundary contacts the core boundary at some temperature (~4,750 K in Fig. 7.18, but this temperature is uncertain due to uncertainties in the position and slope of the boundary). The Earth's interior was hotter when it first formed, so there would have been no post-perovskite phase. This may have a significant effect on the Earth's evolution, because stability calculations indicate that convection is amplified by a perovskite-post-perovskite phase boundary (see Sect. 7.8). This amplification does not occur if the post-perovskite phase is absent. Therefore, there may have been a relatively sudden increase in mantle convection and the rate of heat flow at some point, perhaps fairly long after the Earth formed.

There could also be implications for mantle convection in smaller planets such as Mercury and Mars, in which pressures are too low for post-perovskite to appear, and more massive extrasolar planets (super-Earths), where the post-perovskite layer would be much more extensive.

In Chap. 8, we will apply this information to material collected from our closest external body, the Moon; and, in Chap. 9, to Mercury, Venus and Mars.

References

Anderson, D.L.: Theory of the Earth. Blackwell, Oxford (1989)

Bassett, W.A.: The diamond cell and the nature of the earth's mantle. Annu. Rev. Earth Planet. Sci. **7**, 357–384 (1979)

Breuer, D., Zhou, H., Yuen, D.A., Spohn, T.: Phase transitions in the Martian mantle: implications for the planet's volcanic evolution. J. Geophys. Res. **101**, 7531–7542 (1996)

Catalli, K., Shim, S.-H., Prakapenka, V.: Thickness and Clapeyron slope of the post-perovskite boundary. Nature **462**, 782–785 (2009)

Dziewonski, A.M., Anderson, D.L.: Preliminary reference earth model. Phys. Earth Planet.In. **25**, 297–356 (1981)

Hirose, K., Lay, T.: Discovery of post-perovskite and new views on the Core–Mantle boundary region. Elements **4**(3), 183–189 (2008)

Liu, L.G.: Phase transformations and the construction of the Deep Mantle. In: McElhinny, M.W. (ed.) The Earth: Its Origin, Structure and Evolution, pp. 117–202. Academic Press, London (1979)

Murakami, M., Hirose, K., Kawamura, K., Sata, N., Ohishi, Y.: Post-perovskite phase transition in $MgSiO_3$. Science. **304**, 855–858 (2004)

Oganov, A.R., Ono, S.: Theoretical and experimental evidence for a post-perovskite phase of $MgSiO_3$ in Earth's D″ layer. Nature **430**, 445–448 (2004)

Panning, M., Romanowicz, B.: Inferences on flow at the base of Earth's mantle based on seismic anisotropy. Science **302**, 351–353 (2004)

Ringwood, A.E., Major, A.: Synthesis of majorite and other high pressure garnets and perovskites. Earth Planet. Sci. Lett. **12**, 411–418 (1971)

Schubert, G., Yuen, D.A., Turcotte, D.L.: Role of phase transitions in a dynamic mantle. Geophys. J. Roy Astron Soc **42**, 705–735 (1975)

Chapter 8
The Moon's Surface, Structure, and Evolution

Galileo provided the first telescopic description of the Moon. Many subsequent atlases have been created, and lunar orbiters have provided detailed images of nearly all of the Moon's surface.

We list the bulk properties of the Moon in Tables 13.1 and 13.2 of Milone and Wilson (2014, Chap. 13) in the context of the other moons of the solar system, but it is relevant to mention some of its unique characteristics here. The Moon is the largest satellite in the inner solar system, with a radius of 1,737 km $= 0.2724\ R_{\oplus}$ and mass $0.0123\ M_{\oplus}$. It is the only extraterrestrial body on which people have actually landed. It has also a unique dynamical history, which we will discuss later in this chapter. First, however, we will discuss its composition and appearance.

8.1 Surface Composition

Sources of information about the composition of the Moon include:

1. Ground-based and spacecraft detectors
2. Lunar lander experiments and retrievals

From (1), we gain details from several types of spectrographs and detectors which were targeted at specific types of phenomena: x-ray fluorescence and gamma-ray spectroscopy, and multispectral imaging in the UV, visible and IR. Ground-based instruments can reveal the presence of minerals (see Chap. 7), primarily through their infrared signatures.

E.F. Milone and W.J.F. Wilson, *Solar System Astrophysics: Background Science and the Inner Solar System*, Astronomy and Astrophysics Library,
DOI 10.1007/978-1-4614-8848-4_8, © Springer Science+Business Media New York 2014

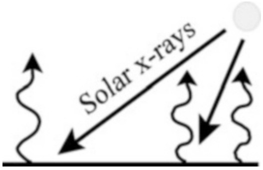


Fig. 8.1 Solar x-rays and fluorescence

8.1.1 *Lunar Orbiting Spacecraft Detections*

8.1.1.1 X-Ray Fluorescence

The lunar surface materials absorb solar x-rays. Aluminum (Al), silicon (Si), and magnesium (Mg) in the surface layer then fluoresce, a process that involves the cascading of electrons from higher to lower energy levels, to produce longer wavelength radiation (Fig. 8.1). Al, Si, and Mg abundances in the lunar surface materials have been measured to a resolution of about 20 km.

Observed ratios of these elements allow a determination of certain major mineral types in the lunar surface layer. Two important examples, their chemical makeup, and main characteristics are:

1. Anorthositic plagioclase feldspar:
   $CaAl_2Si_2O_8$—high Al/Si ratio, low Mg/Si ratio
2. Lunar basalt: pyroxene ($MgSiO_3$) + feldspar—low Al/Si ratio, high Mg/Si ratio

8.1.1.2 Gamma-Ray Spectrometry

Gamma rays (frequently written γ-rays) are high-energy photons emitted by the lunar surface. There are two components, arising from different sources:

1. γ-rays emitted by the radioactive decay of uranium (U), thorium (Th), and potassium (K) and their daughter products in the top few centimeters of the lunar regolith
2. Emission resulting from irradiation of the regolith by galactic cosmic ray particles

   Detectors in the spectral region 0.54–2.7 MeV reveal mainly the component due to radioactive decay.

8.1.1.3 Multispectral Imaging

The spectrum of sunlight reflected from a rocky surface such as the Moon or Mars is different for rocks of different mineral compositions; therefore, brightness differences and brightness ratios between various wavelengths in the IR, visible and UV can be used to measure abundances of minerals such as orthopyroxene, clinopyroxene, and olivine, and elemental abundances of metals bound as oxides in silicates, such as FeO, TiO_2 and Al_2O_3. Multispectral imaging has been done using ground-

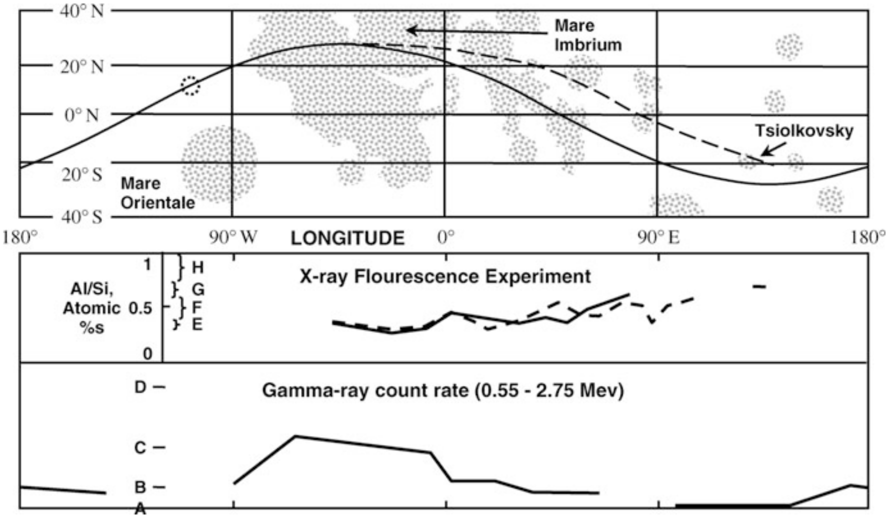


Fig. 8.2 Large-scale compositional variation on the Moon as determined by *Apollo 15* orbital experiments. The *solid* and *dashed lines* in the lower two plots (depicting the Al/Si ratios from the x-ray fluorescence experiment and the γ-ray counts, respectively) refer to data taken along the *solid* and *dashed trajectories* in the map. The γ-ray counts are incompletely reduced here, but include data from earlier *Apollo* missions. Symbols: A: inert material; B: *Apollo 11* soil; C: *Apollo 12* soil 12070; D: *Apollo 14* soil; E: mare basalts; F: lunar norites; G: anorthositic gabbros; H: gabbroic anorthosites. The *dotted circle* on the map shows the approximate spatial resolution of the instruments. From Wood (1972) for the Lunar Sample Analysis Planning Team and adapted with permission from the AAAS. Gabbro is the plutonic form of basalt, with an equivalent chemical composition but solidifying at depth. When anorthositic plagioclase crystallizes from a melt, it can trap residual, mafic liquid that subsequently solidifies to form gabbro. The accepted terminology for lunar rocks is (Stoffler et al. 1980): anorthosite contains <10 % mafic minerals (orthopyroxene, clinopyroxene and/or olivine); gabbroic anorthosite is between 10 % and 22.5 % mafic; anorthositic gabbro is between 22.5 % and 40 % mafic; and gabbro is >40 % mafic—the rest being plagioclase (primarily anorthosite) in all of these cases

based telescopes, and cameras on spacecraft; e.g., the *Clementine* orbiter (1994) mapped abundances over almost the entire lunar surface with an average resolution of 200 m, and the Moon Mineralogy Mapper on the *Chandraya'an* orbiter (2008–2009) mapped much of the lunar surface at a resolution of ~140 m.

8.1.2 Landers (1969–1973)

From the *Apollo 11–17* (excluding *13*) missions, 382 kg of samples were returned. Heat flow was measured and seismic stations were set up. Three Soviet unmanned *Luna* landers were also deployed and a total of 0.326 kg of samples was returned. The *Apollo* command modules contained detectors that were able to survey the Moon. Figure 8.2 summarizes the surveys' highlights.

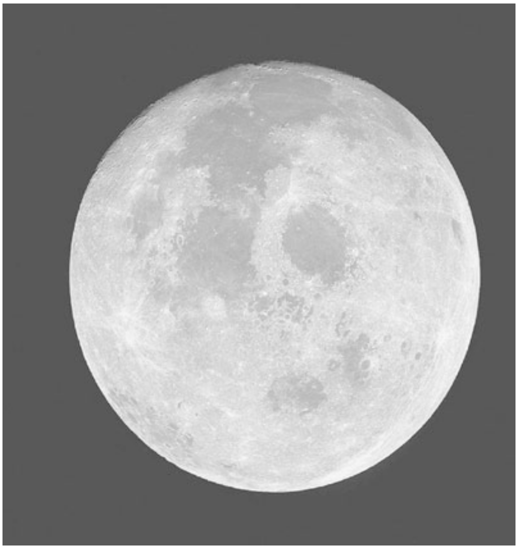


Fig. 8.3 *Apollo 11*'s crew caught this view of the Moon in which the familiar eastern edge is down and near the center of this image, whereas the north limb is to the upper right (compare to the near- and far-side facings in Figures 8.4 and 8.5, respectively). The circular basin to the right of center is Mare Crisium, and that to the left of center is Mare Fecunditatis. Mare Tranquillitatis is above it, and near the upper limb is Mare Serenitatis. The lunar Pyrenees separate Mare Fecunditatis from Mare Nectaris on the left, above the bright-rayed craters. NASA photo AS11-44-6665

The brightness variation on the Moon can be viewed in Figs. 8.3, 8.4, and 8.5 and are described in the next section.

8.2 Lunar Surface Characteristics

The traditional division of the lunar surface is into two parts, apparent in Figs. 8.3, 8.4, and 8.5:

- Bright highlands (or *terrae*; light, rough rock), making up more than 80 % of the lunar surface.
- Dark maria (singular: *mare*, meaning "sea"): flat, relatively smooth lava plains composing ~16 % of the lunar surface.

The maria are unevenly distributed: they cover almost 32 % of the near side, but only about 1 % of the far side. Many are round, lava-filled impact basins (e.g., Mare Imbrium), but there are other large, lava plains for which no impact basin can be discerned (e.g., Oceanus Procellarum). Moreover, the South Pole-

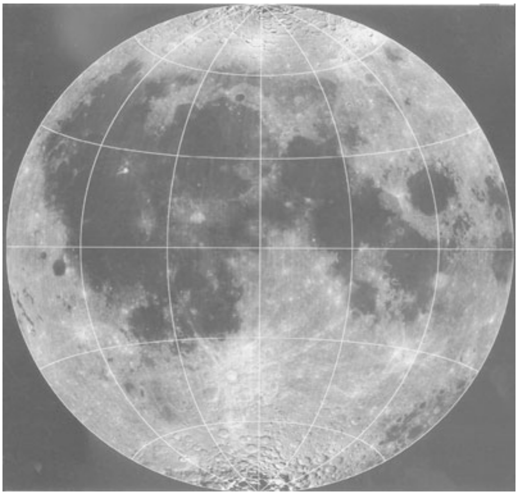


Fig. 8.4 The US Geological Survey albedo (relative brightness) map of the near side of the Moon produced with data from the *Clementine* lunar mission at a wavelength of 0.75 μm. North is up and East is to the right. Major and minor circle arcs mark 30º intervals. The prominent bright-rayed crater just west of the central longitude is Tycho, and the prominent dark-floored crater much farther north is Plato. Mare Crisium is to the right. Note the darkness of the maria compared to the highlands. Reproduced, courtesy, US Naval Research Laboratory and US Geological Survey

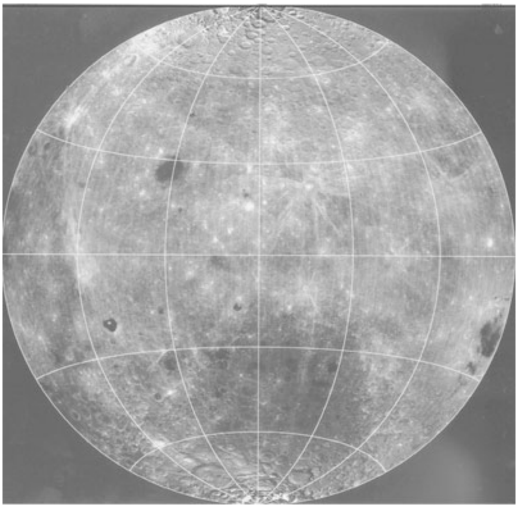


Fig. 8.5 The US Geological Survey (USGS) albedo (relative brightness) map of the far side of the Moon produced with data from the *Clementine* lunar mission at a wavelength of 0.75 μm. North is up and East is to the right. Major and minor circle arcs mark 30° intervals. Note the relative paucity of maria. The dark crater with brilliant white peak, located at ~50°W longitude and —20°S, is Tsiolkovsky. Reproduced, courtesy, US Naval Research Laboratory and US Geological Survey

Aitken basin, located almost entirely on the lunar far side, is the largest impact feature so far recognized in the solar system, yet in contrast to the large, near-side basins it contains very little lava.

Next we summarize the main types of surface materials.

8.2.1 Regolith

The lunar surface has been pulverized by meteoroid and micrometeoroid impacts (the latter being <1 mm in size), and further weathered by the solar wind, particle radiation from solar flares, and cosmic rays. The evidence is to be found in the powdery layer of pulverized rock known as *regolith.* (The term regolith refers to all crushed materials on the lunar surface; the term *lunar soil* refers to the component of the regolith with particle sizes <1 cm.) The depth of the regolith varies over the lunar surface; e.g., at three sites in Oceanus Procellarum the depth of the regolith was found to vary between 8 and 35 m even over short distances; e.g., sites only 0.5 km apart (Wilcox et al. 2005).

In the highlands, the regolith is underlain by a *megaregolith*, a porous layer 1–3 km thick of rock fragments ejected by crater- and basin-forming impacts. The megaregolith is also sometimes defined to include the impact-fractured layer below that, which can be a few 10s of km thick.

8.2.2 Breccias

Generally, *breccia* is rock formed when rock fragments are welded together by metamorphism (see Sects. 7.1.3 and 7.1.4). Most lunar rocks are, in fact, breccias:

- Regolith welded together by heating during impacts
- Rock fractured by impacts and welded together by impacts

Breccias often contain fragments of older breccias; some lunar rocks show as many as four generations of breccias. Some meteorites, too, are brecciated (Milone and Wilson 2014, Sect. 15.3.1).

8.2.3 Impact Melts

These are fine-grained crystalline rocks created during impacts, which generate enough heat to melt some of the rock at the impact site. All nonbrecciated rocks found on the lunar surface are solidified impact melts. They are *not* ancient rocks which have escaped impact brecciation.

8.2.4 Transient Lunar Phenomena (TLPs)

Transient phenomena are occasionally reported at or just above the lunar surface by Earth-based observers. Event durations range from tenths of a second to several hours, and can involve temporary brightening, dimming or obscuration, and color changes. Most have been observed visually through a telescope, with no permanent record such as a video or confirmation by an independent observer; therefore, their reality has been disputed. Crotts (2008) has analyzed statistically a subset of these events, "subjected to a discriminating statistical filter robust against sites of spurious reports," and found this subset to be highly correlated with the mare-highland boundary. Deep and shallow moonquakes also show a correlation of this nature, as does outgassing of radioactive ^{222}Rn and ^{210}Po (detected by alpha particle detectors on Lunar Prospector and the Apollo orbiters). Some TLPs may thus result from gas released by the gradual sagging of mare basalt plains. A lower but possibly significant correlation exists between TLPs and young (<~0.5 Gy) craters such as Aristarchus and Kepler, where fractures or perhaps avalanches could perhaps release gas.

Another kind of transient lunar phenomenon has been well-attested to by both amateur and professional astronomers monitoring the Moon with video cameras—meteoroid strikes on the dark side of the Moon. With the possibility of a manned return to the Moon sometime in the future, observations like this are important for determining the risk of meteorite impacts to astronauts on the lunar surface.

8.2.5 Terrae (Highlands)

The surface has been totally pulverized by impacts. As a consequence, there are no bedrock exposures.

8.2.5.1 Ages of Highland Rocks

From radioisotope measurements (see Milone and Wilson 2014, Sect. 15.5.1), the ages of lunar rocks are found to be mostly 4.0–3.8 Gy, due to "resetting" of many age clocks by impacts; but some rocks have ages approaching 4.5 Gy, and one sample of dunite (almost pure olivine) has an age of 4.55 Gy.

The interval 4.0–3.8 Gy is brief in geological terms. Is it due to a short-lived high-impact era? Or does it represent simply the tail end of a protracted impact era extending from the accretion of the moon?

8.2.5.2 Dominant Highland Mineral

One mineral has a commanding presence on the lunar surface:

Anorthositic plagioclase feldspar, $CaAl_2Si_2O_8$ (see Sect. 7.2.8). This is generally 90–98 % anorthite (Ca, Al-rich). Note that the name *plagioclase* covers a complete series: $NaAlSi_3O_8$–$CaAl_2Si_2O_8$, from *albite,* $NaAlSi_3O_8$, to *anorthite,* $CaAl_2Si_2O_8$; in between are *oligoclase, andesine, labradorite,* and *bytownite.*

Lunar rocks also contain up to 10 times more titanium than terrestrial rocks. *Ilmenite* ($FeTiO_3$) is a common mineral containing this element. These data are important clues to the origin and evolution of the Moon.

8.2.5.3 Highland and Other Non-Mare Rock Types

Most, but not all, rocks that do not originate in mare lava fields come from the highlands. These non-mare rocks are generally grouped into three different suites (Wieczorek *et al*. 2006; Warren 2003):

1. Ferroan anorthosite suite (FA suite or FAS) : Typically >94 % anorthositic plagioclase ($CaAl_2Si_2O_8$) with only ~1–3 % by volume of mafic minerals (pyroxenes, olivine). The pyroxenes can be either:

   low-Ca pyroxene: typically,

   $MgSiO_3$ Enstatite 52–67 molar % (per cent abundance by mole)
   $FeSiO_3$ Ferrosillite 30–50 molar %
   $CaSiO_3$ Wollastonite 1.5–3 molar %

   high-Ca pyroxenes:

   $MgSiO_3$ Enstatite 30–45 molar %
   $FeSiO_3$ Ferrosillite 12–29 molar %
   $CaSiO_3$ Wollastonite 40–46 molar %

   Lower Mg, higher Fe content than the Mg-suite (see below):

   Mg#[1] $\equiv$ mol % Mg/(Mg + Fe) = 0.43–0.75

   Low KREEP and thorium content (Sect. 8.2.7).

   Ages >4.5 Gy–4.3 Gy. The FA suite rocks are believed to have originated during crystallization of a lunar magma ocean (Sect. 8.6). In this case their ages should all be close to 4.5 Gy when this ocean was molten, so it is possible that the apparently-younger rocks have been disturbed or reset in some way.

2. Magnesian suite (Mg-suite):
   Significantly less common than FA suite rocks; ages >4.5–4.1 Gy.
   Major minerals: Plagioclase (~35–85 %), low-Ca pyroxene, and olivine.
   Generally higher Mg, lower iron content than FA suite: Mg# = 0.6–0.95.

[1] Mg# is the ratio of the number of moles of Mg in a sample to the sum of the number of moles of Mg plus the number of moles of Fe. A lower Mg# corresponds to a lower Mg content relative to Fe.

High KREEP and thorium content (Sect. 8.2.7). All Mg-suite rocks are associated with the Procellarum KREEP Terrane (Sect. 8.3, below).

3. Alkali suite:
   Least common and youngest of the three suites; ages 4.3–3.8 Gy.
   Richer in alkali elements (Na, K, Rb, Cs) than other non-mare rocks.
   Wide range of Mg#: <0.05 to >0.90.
   High KREEP and thorium content (Sect. 8.2.7, below). All alkali suite rocks are associated with the Procellarum KREEP Terrane (Sect. 8.3, below).

Both the Mg-suite and the alkali suite appear to have formed from KREEP magmatism after the formation of the lunar crust; although the oldest Mg-suite ages are very similar to those of the FA suite.

8.2.6 Maria

The maria are basaltic lava plains. They are very thin:

A few hundred meters thick on average
2–4 km thick in the centers of the big impact basins

The contrast between a smooth mare surface and rough highland surface is seen in Fig. 8.6, and in a close-up view of the area indicated by the arrow, in Fig. 8.7.

8.2.6.1 Mineral Composition of Mare Basalts

35–68 % clinopyroxene ($CaMgSi_2O_6$–$CaFeSi_2O6$ series)
10–40 % anorthositic (Ca, Al-rich) plagioclase feldspar:
the plagioclase is 60–98 % anorthosite, $CaAl_2Si_2O_8$
0–20 % olivine

The albedos of mare basalts are low. They are dark compared to the highlands (about half the albedo of the highlands) because:

They have higher iron content (due to greater pyroxene/olivine content)

Impact melting produces glass spherules as part of the regolith; glass is darker when the iron content is higher

8.2.6.2 Ages of the Maria

The impact events occurred during the period 4.3–3.8 Gy b.p. (*before present*). Most of these occurred in the period 3.95–3.85 Gy b.p., but the oldest impact basin, South Pole-Aitken (located mostly on the lunar far side, with only one edge visible from Earth), appears to have an age near 4.25 Gy from crater counts (Hiesinger *et al.* 2012) and radiometric dating of some lunar farside meteorites and *Apollo*

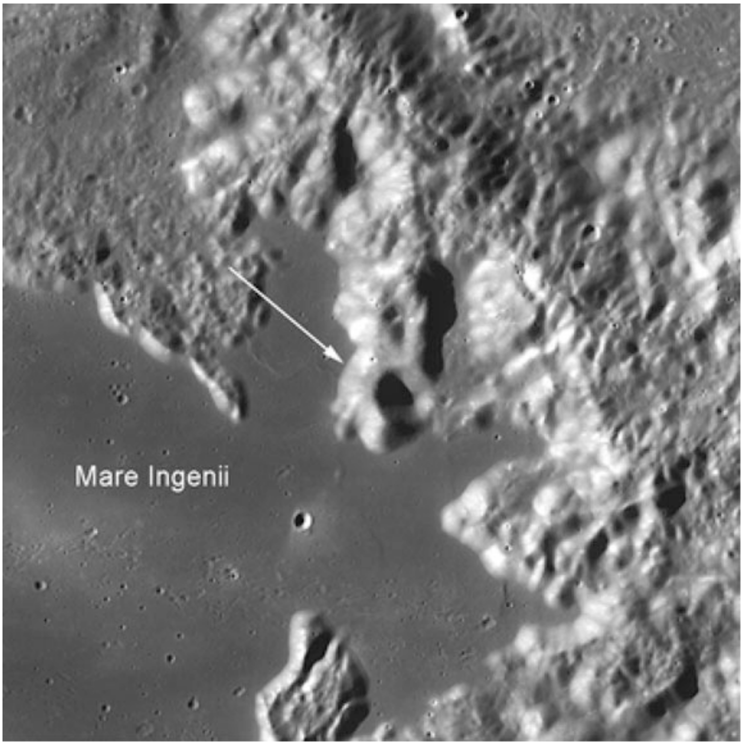


Fig. 8.6 A Lunar Reconnaissance Orbiter Camera (LROC) Wide-Angle Camera (WAC) image of the far side of the Moon, showing the boundary between Mare Ingenii and the surrounding lunar highlands. The lava flows of Mare Ingenii cover an area of diameter ~320 km, of which a 100-km width is shown in the image. The flooded floor of Thompson crater (112 km diameter), located within this mare, is visible in the lower left of the image. The arrow points to the location of the close-up image in Figure 8.7. LROC WAC image M1113062041RE. Credit: NASA/GSFC/Arizona State University

samples that have been interpreted as being from the South Pole-Aitken impact (Garrick-Bethell et al. 2008).

The lava flows filling the nearside impact basins have ages 4.0 to 1.2 Gy from radiometric dating of *Apollo* samples (but only some maria were sampled) and crater counts (Hiesinger et al. 2003). By far the largest proportion of the lava flows occurred in the much narrower period from 3.8 to 3.3 Gy b.p. (Hiesinger et al. 2003 and their Fig. 11).

8.2.6.3 Orange Glass

The *Apollo 17* astronauts found small glass spherules, about the size of silt particles on the Earth, colored orange due to their high titanium content. They are likely to have originated in lava fountains, as lava droplets that solidified in flight. Some were coated with volatiles, e.g., Zn, Pb, S, Cl.

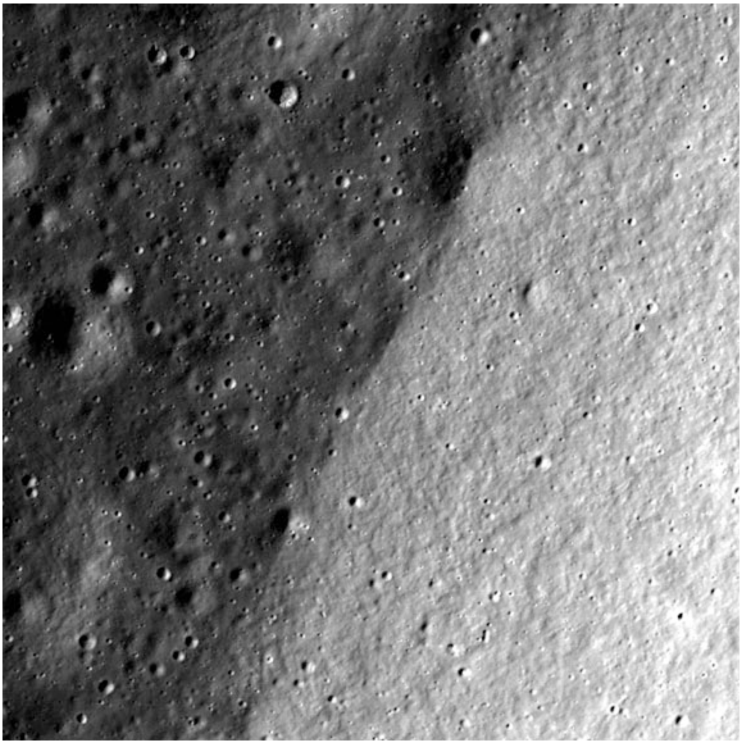


Fig. 8.7 A Lunar Reconnaissance Orbiter Camera (LROC) Narrow-Angle Camera (NAC) image of the far side of the Moon, showing a detailed view of the boundary between the dark, basalt surface of Mare Ingenii on the left and the lunar highlands on the right. The image is 700 m in width, located at the point of the arrow in Figure 8.6. LROC NAC image M1113062041RE. Credit: NASA/GSFC/Arizona State University

8.2.7 KREEP

A proportion of rock samples from all of the *Apollo* landing sites is enriched in, or contains a component that is enriched in, *incompatible elements*; i.e., elements that do not enter minerals easily (Sect. 7.3). These include potassium (K), rare earth elements (REE), phosphorus (P) (hence the acronym, KREEP), thorium and uranium. An earlier name, Low-K Fra Mauro basalt, or LKFM, has been applied to rock of this type, but was never defined clearly and has tended to drop out of use (Korotev 2000).

As an example, the Mare Tranquilitatis regolith sampled by the *Apollo 11* astronauts consists of, by mass (Korotev and Gillis 2001),

66 % crystalline mare basalt
5 % orange volcanic glass

20 % feldspathic highlands material
8 % KREEP-bearing impact-melt breccias
1 % meteoritic material.

The KREEP-bearing breccias in turn have a range of compositions; taken over the samples from *Apollo 14*, *15*, *16* and *17*, the components are (Korotev 2000),

30–95 % KREEP norite (mean 58 %), non-clastic (i.e., it is incorporated into the melt)
1–27 % Mg-rich dunite (mean 13 %), 90 % forsterite, non-clastic
4–50 % feldspathic upper crust (mean 29 %), often clastic (i.e., appearing as lithic inclusions)
0.1–1.7 % Fe/Ni (meteoritic)

(Norite is a plutonic rock similar to basalt.) In addition to the KREEP component in impact-melt breccias, there are also examples of KREEP basalt found as rock fragments in the regolith and lithic clasts in breccias.

Clementine data show that KREEP-bearing materials are most common in the Mare Imbrium—Oceanus Procellarum region, where they occur primarily at low elevations, and are rare in the lunar highlands. However, KREEP basalts differ from mare basalts in having generally higher Al_2O_3 content and lower Ca/Al ratios; i.e., higher plagioclase and lower clinopyroxene content. Evidently, KREEP is neither highland nor mare in origin. The usual interpretation is that it is lower-crustal material (with the dunite possibly being upper mantle) thrown out by the Imbrium and perhaps Serenitatis impacts. The fact that the KREEP and dunite are mixed in the melt and the feldspathic upper crust occurs largely as clasts suggests that the impact site may have had little or no feldspathic crust, and the crustal clasts were incorporated into the ejected melt near or beyond the edge of the impact basin (Korotev 2000). Thus, the Imbrium-Procellarum region seems to have differed from the highland areas even before the impact(s) occurred.

In Sect. 8.3, we look at some consequences of these observations.

8.3 Lunar Terranes

Results from *Clementine* (1994), *Lunar Prospector* (1998–1999), and subsequent missions indicate that the lunar surface is more complex than can be accommodated by a simple two-part division into highlands and maria; e.g., the KREEP component discussed in the previous section. In its place, Jolliff *et al.* (2000) propose that the crust and upper mantle consist of at least three different geologic *terranes* of differing character (see also the review by Wieczorek *et al.* 2006). Unlike the traditional division, in which the maria can be regarded simply as surface modifications by impacts and lava flows on a globally-uniform highland crust, terranes represent geologically-distinct regions that differ from each other at depth (crust and possibly the underlying upper mantle) as well as at the surface.

The three main terranes, with some subdivisions and the fraction of the lunar surface between 60° N and 60° S that each terrane occupies, are

Procellarum KREEP Terrane (PKT): 16.5 %

An approximately oval area containing Oceanus Procellarum and related maria and Mare Imbrium, with some adjacent non-mare terrain. The PKT is defined primarily as the region for which thorium concentrations are >3.5 ppm.

South Pole-Aitken Terrane (SPAT): 11.0 %

Occupies most of the farside southern hemisphere. It has subdivisions:
SPAT-inner—the primary South Pole-Aitken impact basin.
SPAT-outer—the rim and adjacent ejecta blanket.

Feldspathic Highlands Terrane (FHT): 59.5 %

FHT-Anorthositic (FHT-A)—the area of thickest crust, occupying most of the farside north of the SPAT and including the most strongly anorthositic regions of the lunar surface (80–90 % plagioclase).
FHT-outer—everything surrounding FHT-inner, on both the near and farside, that is not PKT, SPAT or "other mare" (see below). It consists of the older, presumably feldspathic surface that has been obscured by basin ejecta.

Other, smaller areas that are not part of the above three terranes are collectively referred to as "other mare (OM)," including Mare Serenetatis, Tranquillitatis, Crisium and Australe, total 13 %.

In addition to mineral content, another useful way to classify rocks is by their abundance of certain elements, notably iron, titanium and thorium. Where the element is bound with oxygen in minerals (e.g., $FeSiO_3$ (fayallite), $FeTiO_3$ (ilmenite)), it is conventional to specify the element in terms of its oxidation state; e.g., in $FeTiO_3$, the valences are Fe^{2+}, Ti^{4+}, and O^{2-}, so one O^{2-} can be associated with the Fe^{2+} and two with the Ti^{4+}. Thus, when specifying the abundances of iron and titanium bound within minerals, $FeTiO_3$ contributes to both FeO and TiO_2. Thorium is of interest because it is a heat-producing element through radioactive decay. Being an incompatible element, it also tends to become concentrated in the residual melt as other phases crystallize in a solidifying magma (Sect. 7.3).

The iron and thorium abundances in Table 8.1, adapted from Jolliff et al. (2000), illustrate the differences in character of the three terranes.

From Table 8.1, thorium tends to be concentrated in a single, well-defined area (the PKT) taking up a relatively small fraction (16.5 %) of the lunar surface. This suggests fundamental differences in composition of the underlying crustal or upper mantle material when this terrane formed.

From *Apollo* samples, FeO correlates inversely with Al_2O_3, which is a component of plagioclase ($CaAl_2Si_2O_8$). The low FeO abundance of 4.2 wt% (read: percent by weight) in FHT-A corresponds to a high Al_2O_3 abundance of 26 wt%, making FHT-A highly anorthositic; its average composition is that of noritic

Table 8.1 Lunar Terranes

	FeO (wt%)	Th (ppm)
FHT-Anorthositic	4.2	0.8
FHT-Outer	5.5	1.5
Other Mare	16.2	2.2
OM, mixed[a]	8.8	1.6
PKT-nonmare	9.0	5.2
PKT-mare	17.3	4.9
PKT-mixed	10.7	4.5
SPAT-inner	10.1	1.9
SPAT-outer	5.7	1.0

[a] "Mixed" refers to areas where each pixel includes both mare and nonmare characteristics

anorthosite, which is 80–90 % plagioclase. Even the largest impacts in FHT-A, which are believed to have penetrated up to 25–35 km depth, did not expose material of substantially higher FeO content, indicating that the crust in FHT-A is highly anorthositic to at least this depth.

8.4 Water on the Moon

It has been suggested that permanently-shadowed craters at the lunar poles, where the temperature remains below 50 K, could act as cold-traps, allowing water to exist there in the form of ice. In 2009, NASA tested this idea directly by crashing a spent Centaur rocket stage into a permanently-shadowed part of Cabeus crater, ~8.5° from the lunar south pole (Colaprete *et al.* 2010). The impact and debris plume were closely observed by the *LCROSS* spacecraft (Lunar Crater Observation and Sensing Satellite), which followed the Centaur down to the lunar surface and crashed nearby 4 min later, and from orbit by *Lunar Reconnaissance Orbiter* (*LRO*).

The debris plume from the Centaur impact reached sunlight 1 s after impact, with maximum brightness 17 s after impact, and remained substantially brighter than the shadowed background for the remainder of the 4 min to *LCROSS* impact.

Figure 8.8 shows the near IR spectrum of the vapour and debris plume observed by the *LCROSS* satellite in two separate time intervals. Water ice absorption is clearly visible in the spectrum between 23 and 30 s after impact, with small amounts of SO_2 and water vapour. The water ice feature confirms that the water in the crater was in the form of ice and not simply water adsorbed on grains. Due to continued sublimation of the ice grains and possibly of ice at the hot impact site, the spectrum between 123 and 180 s still shows water ice, but with stronger water vapour absorption, along with some CH_4 and perhaps CO_2. The maximum mass of water observed in the field of view (FOV) of the instrument occurred during the interval 23–30 s after impact: 24.5 ± 8.1 kg vapour + 131 ± 8.3 kg ice = 155.5 ± 16.4 kg. The mass of dust in the same FOV was 2,434 ± 609 kg, giving a

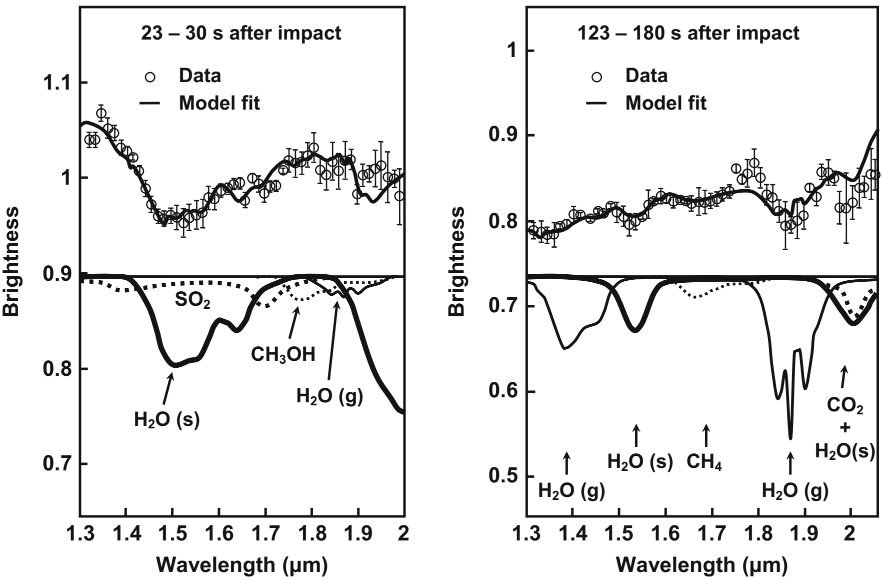


Fig. 8.8 Near IR spectrum of the vapour and debris plume from the Centaur impact, as measured by the *LCROSS* satellite (Colaprete *et al.* 2010, adapted with permission from the authors and publisher). The data points (*circles* with *error bars*) show the brightness averaged over the time period given at the top of each plot. The *solid line* through the data is a model fit produced by combining the spectra of several volatiles, of which the dominant species in the spectral range shown are illustrated in the lower curves. H_2O (g) is water vapour and H_2O (s) is water ice. These spectra have been normalized to that of water ice

fractional abundance of water in the regolith at the impact site of 6.4 ± 1.7 wt% (the fractional uncertainties in the water and dust are added in quadrature).

The source of the water is believed to be comet impacts, and also solar wind interactions with the lunar surface: incoming solar wind protons can liberate oxygen ions from silicates and other minerals, allowing the formation of OH and H_2O. However, evidence has been found for water in the Moon's interior with an abundance similar to that of the Earth's mantle (Sect. 8.7), so indigenous lunar water reaching the surface could also contribute. A possible transport mechanism is suggested by a lunar hydration cycle found by the *Deep Impact* spacecraft as it flew by the Moon once in December, 2007, and twice in June, 2009 (Sunshine *et al.* 2009). IR observations from the spacecraft confirmed the presence of OH and/or H_2O on the dayside lunar surface, with maximum hydration close to the sunrise and sunset terminators. The implied cyclic loss and recovery suggest that hydration is maximum overnight, with H_2O being photodissociated to OH and H^+ after sunrise and recombining toward sunset. OH and H^+ finding its way to the poles could then accumulate as water in permanently shaded regions, providing a slow, net poleward transport of water. Liu et al. (2012) have also confirmed the presence of OH in Apollo 11, 16 and 17 soil. The solar wind was confirmed as the primary source of the hydrogen by the D/H isotope ratios (D = deuterium).

Solar-wind-implanted OH and H_2O affect only the outer 0.2 μm of an exposed rock surface. Hui et al. (2013) looked for water in mineral samples from the interiors of ferroan anorthositic and Mg-suite rocks, and found it at a level of ~4 ppm. Being from the interior of the rock, these samples had been well protected from the solar wind, indicating that this water was indigenous to the Moon. Geochemical considerations of a crystallizing magma ocean then suggest that the original magma ocean, while fully molten, could have had a water content of 320 ppm. The question of indigenous water in the Moon is discussed further in Sect. 8.7.

8.5 Internal Structure

A number of methods have been used to probe the internal structure of the Moon, including

- Seismometers placed on the lunar surface by the *Apollo 12*, *14*, *15* and *16* astronauts.
- Determination of the moments of inertia of the Moon from how it perturbs the motion of orbiting spacecraft. The moments of inertia provide information about the mass distribution within the Moon.
- Measurement of the induced magnetic dipole moment of the Moon by magnetometers in orbit and on the lunar surface. This provides information about electrical conductivity as a function of depth, which in turn provides constraints on composition, including the presence or absence of an electrically-conducting core.
- Lunar laser ranging from Earth using retroreflectors on the lunar surface at the *Apollo 11*, *14* and *15* landing sites and on the two Soviet rovers, *Lunakhod 1* and *2*. The motion of the Moon as given by these retroreflectors provides information about how the Moon's rotation varies with time due to the gravitational influences of the Earth and Sun. These variations in turn depend on the internal structure, including the presence or absence of a molten core.
- Measurement of the tidal distortion of the Moon by the Earth, which provides information about the elastic properties of the lunar interior. Usually, the tidal response is characterized by the potential Love numbers (Sect. 5.3). Of these, the tidal Love number k_2 is particularly useful, describing the changes in the Moon's gravitational potential in response to the Earth's gravitational potential, as described in Sect. 5.3.
- Gravity and topography. The gravitational acceleration of an orbiting spacecraft varies with time because the mass distribution below the spacecraft changes as the spacecraft moves over the surface. After subtracting the variations expected from the observed topography, the remaining variations must be due to the internal mass distribution.
- Thermal and geochemical constraints.
- Studies of paleomagnetism to look for evidence of an ancient global magnetic field. The Moon currently has no global magnetic field, but if one existed in the

ancient past, it would show that the Moon had a molten, electrically-conducting core at that time.

See the review of lunar structure by Wieczorek *et al.* (2006) for more information on many of these methods.

The *Apollo* seismometers were placed on the lunar surface in the period 1969–1972, and provided data until funding was removed in 1977. More than 12,000 events were detected, but the data have been difficult to interpret because the anhydrous nature of the lunar interior (lacking water) causes seismic waves to reflect back and forth with very little attenuation, sometimes for hours, making the seismograms very noisy. Waves arriving directly from the source (an impact or moonquake) can usually be identified unambiguously, but the much weaker waves that are reflected from boundaries within the Moon are very difficult to distinguish from the seismic noise. It is only in the last decade or so that modern numerical techniques have allowed this information to be extracted with reasonable confidence; e.g., improved methods of stacking many seismic records with suitable time delays that depend on the locations of the sources and receivers (Gagnepain-Beyneix *et al.* 2006; Garcia *et al.* 2011; Weber *et al.* 2011).

Seismic events on the Moon fall into four categories:

1. Deep quakes:
   - The most common type of moonquake.
   - Occur in localized regions ("nests") at depths between ~700 and ~1,150 km, with a broad peak between ~900 and ~1,000 km.
   - Relatively weak: <3 on the Richter scale.
   - Tend to repeat in a 27.5-day pattern (the Moon's orbital period around the Earth), so apparently caused by tidal effects.
2. Thermal quakes:
   - The second most common type.
   - Occur at depths within a few km of the surface.
   - Tend to repeat in a 29.5-day pattern (the length of a solar day on the Moon), with a sharp increase in number of events ~48^{h} after each sunrise.
   - Caused by thermal stresses along fracture planes as the sun warms the crust after 2 weeks of darkness. The temperature difference between night and day on the Moon is more than 200 ° C.
3. Shallow quakes:
   - The least common type: only 28 observed between 1969 and 1977.
   - Occur at depths of 50–220 km.
   - Cause: unknown.
   - Relatively strong: up to 5.5 on the Richter scale.
4. Impacts; about 1,700 recorded.

We discuss the crust, mantle and core separately, below.

8.5.1 The Crust

Both the *Apollo* seismic data and gravity and topography measurements show that the lunar crust is considerably thinner on the nearside than on the farside. A re-analysis of the Apollo seismic data by Chenet et al. (2006) gives crustal thicknesses below the Apollo 12 and 14 landing sites of 33 ± 3 km and 31 ± 7 km, respectively.

The global distribution of crustal thickness has been determined most precisely by the *Gravity Recovery and Interior Laboratory* (GRAIL) mission (Mar.–Dec. 2012), which consisted of two identical spacecraft flying in formation only 175–225 km apart in a nearly-polar orbit only 50 km above the lunar surface. Precise determinations of orbital variations and separation distances between the spacecraft allowed the lunar gravity field to be mapped to ~26-km resolution, more than a fourfold improvement over earlier results.

The average density of the lunar crust from GRAIL is 2,550 ± 18 kg m^{-3} (average uncertainty), with lateral variations of up to 250 kg m^{-3} above and below this value (Wieczorek et al. 2013). The highest densities are associated with the South Pole–Aitken impact basin, which is compatible with the more-mafic composition of these rocks as determined by remote sensing, and the lowest with the regions surrounding the farside Orientale and Moscoviense basins. An average grain density of 2,927 kg m^{-3} (i.e., the density of the rock material itself, excluding empty spaces or pores) was found using Lunar Prospector elemental abundances and an empirical correlation between grain density and composition. The low average density suggests a considerable amount of porosity due to impact-induced fractures and brecciation; if the surface composition is representative of the entire crust, then the *global mean porosity* (equal to $[\rho_{\text{grain}}-\rho_{\text{bulk}}]/\rho_{\text{grain}}$, where ρ_{bulk} is the density including pores) is 12 %, varying laterally within a range from 4 % to 21 %.

Crustal thickness models (Wieczorek et al. 2013) were computed using constraints of (1) either 30 km or 38 km crustal thickness at the Apollo 12 and 14 landing sites and (2) a global minimum crustal thickness <1 km because at least one large basin-forming impact should have excavated through to the mantle (e.g., the 600-km diameter Crisium basin should have been excavated to ~60 km depth if the impact was vertical, which is much greater than the crustal thickness given by the Apollo seismic data). The resulting global average crustal thickness is 34 km assuming 30 km for constraint (1), above, or 43 km assuming 38 km. The minimum thickness (<1 km) occurs beneath the farside Moscoviense basin and, assuming 30 km at the two Apollo sites, the maximum thickness is ~60 km near the lunar equator north of South Pole-Aitken.

Pore closure by viscous deformation is predicted to occur at depths between 40 and 85 km because of increasing temperature with depth below the lunar surface. Where the crust is thinner than this, substantial porosity could extend into the uppermost mantle.

Earlier seismic studies have given indications of vertical structure within the crust, based on reflections of seismic waves, sudden changes in wave speed at particular depths, and/or conversions of S to P or P to S waves (see Sect. 5.4.1).

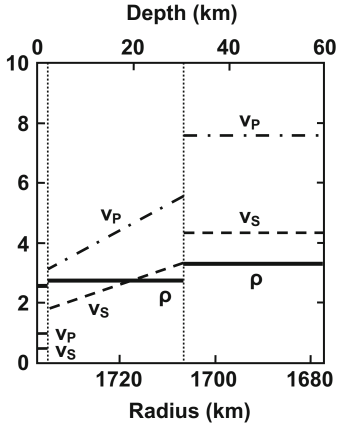


Fig. 8.9 The behavior of V_P (km/s), V_S (km/s) and density (Mg/m^3) in the regolith, crust and uppermost mantle of the Moon, based on Chenet *et al.* (2006) and Gagnepain-Beyneix *et al.* (2006). The left-hand and right-hand vertical, *dotted lines* signify the base of the regolith and the Moho, respectively

Figure 8.9 illustrates the behavior of V_P, V_S (Sect. 5.4.1) and ρ (Sect. 5.4.2) with depth in the area covered by the *Apollo* seismic array:

- 0–1 km depth: Regolith (or perhaps more accurately megaregolith): V_P ~1 km/s; velocity discontinuity to ~3.5 km/s at 1 km depth.
- 1–30 km depth: Continuous increase in V_P from ~3.5 km/s to ~5.6 km/s with no discernable discontinuities or conversions (Gagnepain-Beyneix *et al.* 2006). We illustrate this increase schematically in Figure 8.9 by straight lines, but the uncertainties in V_P (and V_S) allow other interpretations.
- ~30 km depth: S to P wave conversion occurs and V_P increases from ~5.6 km/s to ~7.6 km/s. (The behavior of V_S follows a similar pattern to the above.) The corresponding density increase is 0.55 kg/m^3, from 2,760 kg/m^3 in the crust to 3,310 kg/m^3 in the uppermost mantle (Garcia *et al.* 2011), or from 2,800 kg/m^3 to 3,350 kg/m^3 (Chenet *et al.* 2006). Pressure is too low for a phase change, so the discontinuity is taken as identifying the lunar Moho, or crust-mantle boundary, where the composition changes from anorthositic crust to olivine/pyroxene mantle. Recall from Table 5.1 that the mean uncompressed lunar density is only 3,300 kg m^{-3}.

8.5.2 The Mantle

Two recent models of the lunar interior, from Weber *et al.* (2011) and Garcia *et al.* (2011), are shown in Figs. 8.10 and 8.11. Seismic velocities above ~1,000 km depth were modeled from previous work, and then interfaces below this depth were searched for. The two models for the core region illustrate the similarities and differences in the results obtained by different approaches to analyzing the *Apollo* seismic data.

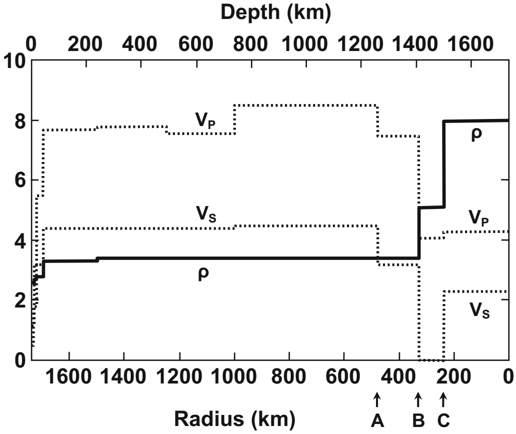


Fig. 8.10 P-wave velocity, V_P (km/s), S-wave velocity, V_S (km/s), and density, ρ (kg/m^3/1,000), through the crust, mantle and core of the Moon from Weber *et al.* (2011, plotted from the values in Table S2 of their online supplement). A: partial-melt boundary (PMB); B: core-mantle boundary (CMB); C: inner-core boundary (ICB). V_S is taken to be zero in the molten outer core (240 km < r < 330 km). Compare Figures 8.9, 8.10, and 8.11 to Figures 5.23–5.25 for the corresponding plots for the Earth's interior

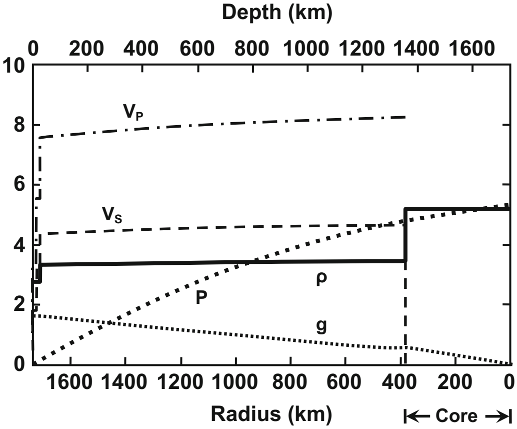


Fig. 8.11 P-wave velocity, V_P (km/s), S-wave velocity, V_S (km/s), density, ρ (kg/m^3/1,000), pressure, P (GPa), and gravitational acceleration, g (m/s^2), through the crust, mantle and core of the Moon as given by the Very Preliminary REference MOON model (VPREMOON) of Garcia *et al.* (2011, plotted from their Table 6, but see their Erratum (2012)). The authors felt that the seismic data did not allow them to identify a solid, inner core, so the core (r < 380 km) is taken to be entirely molten, with $V_S = 0$ and V_P unspecified

Seismic studies from *Apollo* data suggest a division into upper, middle and lower mantle, but the boundaries, seismic structure, composition, and even existence of these regions are still uncertain (Gagnepain-Beyneix *et al.* 2006; see also the review by Wieczorek *et al.* 2006). Gagnepain-Beyneix *et al.* (2006) find the following structure:

Upper mantle (from the Moho to 238 km depth):

Well-determined seismic velocities, $V_P = 7.65 \pm 0.06$ km/s, $V_S = 4.44 \pm 0.04$ km/s; see Fig. 8.10, where the seismic velocities above 480 km radius used by Weber *et al.* (2011) are those of Gagnepain-Beyneix *et al.* (2006), except for a 5 % increase in V_P (8.5 km/s instead of 8.15 km/s) between 738 and 1,238 km depth. The values of Gagnepain-Beyneix *et al.* (2006) are consistent with a pyroxenite upper mantle, while the higher V_P of Weber *et al.* (2011) would require some garnet.

Middle mantle (238–738 km depth):

A low-velocity zone from 438 to 738 km depth (Fig. 8.10), but Weber et al. note that others have found a low-velocity zone in a different depth range, and comment that, given the uncertainties, their own results are consistent with constant velocities in the middle mantle.

Lower mantle (the region of deep moonquakes, below 738 km depth):

$$V_P = 8.15 \pm 0.23\ \mathrm{km/s}\ \text{and}\ V_P = 4.50 \pm 0.10\ \mathrm{km/s}.$$

Two deep moonquake nests are located on the lunar farside (Nakamura 2005), one at 26° ± 9° beyond the eastern limb and the other 21° ± 14° beyond the eastern limb. For both, S waves were not detected at the two seismic stations for which the ray path was deepest, although P waves were detected at one or both stations. These results suggest a zone of partial melt in the lower mantle where S waves are substantially attenuated.

Earlier models had indicated a seismic discontinuity near 550 km depth, with the suggestion that it might represent a composition discontinuity between differentiated cumulates at the base of an ancient lunar magma ocean and an undifferentiated lower mantle; but no discontinuity is apparent at this depth in Gagnepain-Beyneix *et al.* (2006).

8.5.3 The Core

A number of lines of evidence point toward the existence of a molten, metallic core:

1. Magnetic dipole moment:
   When the Moon enters a lobe of the Earth's geomagnetic tail,[2] it encounters an external magnetic field (the magnetotail field) that is almost spatially-uniform

[2] The Earth's magnetosphere and geomagnetic tail are discussed in Chap. 11 of Milone and Wilson (2014).

(strength: ~12–16 nT), and considerably stronger than the field outside the lobe. It is also in a good vacuum, which simplifies magnetic-field interpretations. The Moon is a conducting medium, so when the field external to the Moon changes, this change diffuses into the Moon over time. If we consider a closed path within the Moon, the changing flux within this path induces an emf around the path by Faraday's law, and because the medium is conducting, there is an induced current that (again by Faraday's law) opposes the change in flux; i.e., a diamagnetic current. The current creates an induced magnetic dipole moment that can be measured by magnetometers on orbiting spacecraft. Because the diamagnetic currents oppose the change in flux, the external field change diffuses into the Moon more slowly the greater the conductivity of the medium. The induced current and induced dipole moment are greatest initially and decay exponentially to zero with a time constant that is proportional to the conductivity.

The time for the Moon to cross a lobe of the geomagnetic tail can exceed 2 days. The conductivity of the mantle is ~10^{-1} to 10^{-3} mho/m, giving time constants <1 h, whereas the conductivity of the core is essentially infinite for the timescale of the lobe crossing (Hood *et al.* 1999). Mantle currents then become negligible a few hours after the Moon enters the geomagnetic tail, and any remaining dipole moment is due to the core. Hood *et al.* (1999) measured an induced lunar magnetic dipole moment of $-2.4 \pm 1.6 \times 10^{22}$ Gauss-cm^3 per Gauss of applied field (the negative sign is because of the diamagnetic nature of the current), indicating a highly-conducting core of radius 340 ± 90 km.

Measurement of the dipole moment does not indicate whether the core is solid or molten. This deficiency is addressed in point 2, below.

2. Lunar laser ranging:
The Moon's rotational pole and orbital pole are tilted in opposite directions from the ecliptic pole by 1.54° and 5.145°, respectively, and both precess around the ecliptic pole in a direction counter to the orbital motion with the same period, 18.6 years. If there were no dissipation then the three poles would be coplanar; however, energy dissipation within the Moon causes the spin pole to be advanced by 0.263" in the direction of precession, relative to the orbital pole. The dissipation can be due to solid-body tides raised by the Earth and (to a lesser extent) the Sun, and/or a fluid core whose rotation differs from the solid mantle.

The advance of the spin pole could be caused by either or both of tidal effects and core-mantle interactions, so it does not distinguish between them. However, there are smaller effects (Williams *et al.* 2001) that, with the 28 years of lunar laser ranging data available at that time, are measurable; e.g., the 18.6-year mantle precession induces a core precession with a much smaller tilt. The core rotates at a different rate than the mantle, and core dissipation then causes a measurable shift in the mantle pole.

The smaller terms can be satisfied only if there is both tidal and core dissipation (Williams *et al.* 2001), with a derived radius for the molten core of 352 km if pure iron, or 374 km if it is an Fe-FeS eutectic (Sect. 5.4.4.5).

3. Thermodynamic considerations:
   Thermal evolution models, based on factors such as measured surface heat flow, accretional and radioactive heating, and thermal conductivity of mantle compositions, place the central temperature of the Moon in the range 1,000-1,480 °C, whereas the melting temperature of pure iron at core pressures (~4 GPa) is 1,690 °C (Wieczorek *et al.* 2006). Thus, while a pure iron core could have been molten early in the Moon's life, it would be solid now.

   The molten core indicated by lunar laser ranging therefore suggests the presence of an impurity that lowers the melting temperature, the most likely of which appears to be either carbon or sulfur. The eutectic carbon content is 3.5 wt%, with a eutectic temperature at core pressures of 1,175 °C; so if the impurity is C then the Moon would likely still have a molten outer core. If the impurity is S, the eutectic sulfur content is 25 wt% at core pressures, with a eutectic temperature of 950 °C, and the outer core is certainly still molten. In both cases, if the impurity is below the eutectic composition (e.g., <25 wt% S), then iron is the first component to solidify, and sinks toward the centre to create a solid inner core of pure iron (Wieczorek *et al.* 2006).

4. Seismic reflections:
   Weber *et al.* (2011) and Garcia *et al.* (2011) have re-examined the *Apollo* seismic data using more sensitive methods to reduce noise and allow detection of reflections from deep interfaces. In both cases, the runs of density (ρ), V_P and V_S with depth are modeled from previous work in order to predict arrival times of waves reflected from boundaries at assumed locations, and then these boundaries are shifted to look for actual arrivals in the data.

   Weber *et al.* use P waves and both horizontally- and vertically-polarized S-waves to look for P to P reflections (i.e., an incident P wave reflecting as a P wave), P to S, S to S and S to P. Values for ρ, V_P and V_S are taken from Gagnepain-Beyneix *et al.* (2006) for the crust and mantle above 1,000 km depth, with a 5 % increase in V_P in the lowermost layer, to look for three boundaries: the top of the partial-melt zone in the lowermost mantle (PMB), the core-mantle boundary (CMB), and the inner-core boundary (ICB). The results are shown in Fig. 8.10: $r_{PMB} = 480 \pm 15$ km, $r_{CMB} = 330 \pm 20$ km, $r_{ICB} = 240 \pm 10$ km.

   Garcia *et al.* assume a homogeneous mantle from the crust to the core, and use horizontally-polarized S waves (SH waves) to look for the CMB. S waves cannot propagate in a fluid medium, so if the outer core is fluid then SH waves are totally reflected at the CMB, whereas some of the energy of a vertically-polarized S wave (SV wave) can propagate into the core as a P wave, reducing the reflected energy. However, using only SH waves prevents searching for an inner core, and the assumption of a uniform mantle prevents searching for a partial melt zone. They obtain ρ(r) by integrating the Adams-Williamson equation (Sect. 5.4.2), and then assume monotonically-increasing V_S and V_P laws consistent with previous work. The results are shown in Fig. 8.11, with $r_{CMB} = 380 \pm 40$ km radius.

8.6 Geochemical Evolution

The discovery of a highly-anorthositic crust (~75 wt% plagioclase) ubiquitous over the lunar highlands led immediately to the idea that at least the outer part, and perhaps all, of the Moon was molten when it first formed. The dominant arguments are (further discussion can be found in Warren 2003),

1. The present highland crust could have formed in only one of three ways: (1) The Moon, or its outer part, was always solid; but then it is difficult to see how such high plagioclase material would collect on the surface. (2) Lava flows onto a previous, solid surface. However, basaltic lava from partial melts in the deep interior of a solid Moon is limited by phase equilibria to ~55 wt% plagioclase, which is much smaller than observed. (3) Crystallization in a magma ocean. Plagioclase is light, and would naturally rise to form a plagioclase-rich crust.
2. The simplest explanation for the distinctively-low Mg# values for FA suite rocks (Sect. 8.2.5.3) is formation in an ultramafic magma ocean with high initial Mg# values. High-Mg olivine and pyroxene crystallize first and sink, leaving the remaining melt with lower and lower Mg content relative to Fe. Plagioclase crystallizes later, and, being lighter, rises to form a crust with low Mg# values, as observed. The low-Mg# mafic minerals that should form at the same time would sink and are not seen at the surface.

The depth of the magma ocean is still very uncertain; models have been proposed with anywhere from ~250 km depth to the entire Moon being molten. It is also unclear whether the magma ocean was non-convective or initially turbulently convective. Crystallization has a different effect in the two cases (Sect. 5.4.4.5):

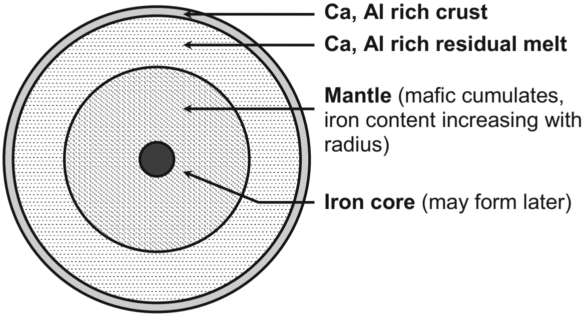
Equilibrium crystallization: If the liquid is turbulently convective, then precipitating crystals remain suspended in the liquid, and the magma solidifies at constant composition. This process continues until the concentration of crystals in the magma-crystal mush exceeds some value, often taken as 50 %, after which convection wanes and the suspended crystals sink.

Fractional crystallization: If the liquid is non-convective (either throughout crystallization if there is no convection, or after convection has ceased), then crystals sink onto a cumulate pile. As the Fe content of the melt increases, the iron content of the olivine and pyroxene precipitating from it increases, producing a differentiated cumulate pile (increasing Fe content with height).

Crystallization sequence (Elardo *et al.* 2011):

1. Mg-rich olivine precipitates first, then orthopyroxene. This creates a dunite (olivine) to harzburgite (olivine + pyroxene) lower mantle. The Fe content of the residual melt increases as Mg is removed, so the precipitate becomes progressively enriched in Fe compared to Mg.

Fig. 8.12 Fractionation scenario for the Moon


Ca, Al rich crust
Ca, Al rich residual melt
Mantle (mafic cumulates, iron content increasing with radius)
Iron core (may form later)

2. Anorthitic plagioclase begins to precipitate after ~75 % of the initial magma ocean has solidified. Being less dense than the melt, it floats to the surface to form the primary lunar highlands crust. The complementary mafic, Fe-rich cumulates continue to sink onto the mantle. The fractionation to this point is illustrated in Fig. 8.12.

3. Ilmenite ($FeTiO_3$) precipitates after ~95 % solidification, creating a very dense mantle layer that is both iron- and ilmenite-rich.

4. The final dregs of the residual magma ocean are enriched in incompatible elements that do not enter mineral phases easily, including potassium (K), rare earth elements (REE), phosphorus (P), and thorium. This final liquid is often referred to as urKREEP.

5. The mantle at this point is gravitationally unstable, with denser, iron-rich silicates and ilmenite above lighter, magnesium-rich silicates. It is possible that this leads to mantle overturn and various amounts of mixing and/or relocation of early- and late-stage cumulates.

6. If this overturn occurs, the Mg-rich silicates rising from below could play a role in forming a more Mg-rich lower crust and the Mg-suite rocks at the lunar surface. Lateral and radial differences in mixing can also create the varied source lithologies for the very diverse compositional range of secondary lunar magmatism (e.g., the mare basalts) and the different lunar terranes.

An unknown process appears to have caused the urKREEP to concentrate in one area of the Moon. The enrichment in K, Th and U in this area would have kept the residual melt molten longer, possibly until the Imbrium impact, giving rise to the Procellarum KREEP Terrane.

Another discovery from the GRAIL mission (Sect. 8.5.1) is an extensive system of short-wavelength, linear gravity anomalies distributed approximately isotropically around the Moon (Andrews-Hanna et al. 2013). Individual anomalies are generally quite straight (roughly along arcs of great circles) for up to 500–1,000 km for the longest ones, with a total length of ~10^4 km for the entire system. Observed widths are ~5–25 km, although smaller ones may be hidden below the resolution of the data. The anomalies are believed to be the gravitational expression of intrusive, mafic dikes with the lengths and widths given above, and depths from ~10–15 km below the surface at the top of the dike to ~75–90 km at the bottom. The radial placement of some of the dikes relative to the South Pole-Aitken basin indicates that they were emplaced after the impact (because they were influenced by it), whereas another dike was already there before the Crisium impact, which erased a portion of dike. The dike system therefore appears to have formed within the first ~1 Gy after the formation of the Moon, providing a window into very early processes in the Moon's life.

An isotropically-distributed system of dikes indicates global extensional stresses in the crust, with magma rising into the resulting fractures. The extensional stresses could arise through expansion of the interior or contraction of the lithosphere around an unyielding interior, or some combination of both. One clue is the absence on the Moon of a global system of large thrust faults like those found on Mercury (Sect. 9.1.3.4.3). It is argued that the latter were created by crustal compression as Mercury's interior cooled and contracted under the already-solidified crust. A similar effect should have happened on the Moon, so their absence suggests that the Moon's interior was initially cooler than the outer layers, and warmed up after the crust had solidified. This could happen if the outer layers were heated externally by impacts, creating a global, 200- to 300-km-deep magma ocean. Heat exchange would mean that the interior would heat up as the outer layers cooled, evidently (as shown by the dike system) providing net expansion for the first ~1 Gy and then only modest (<1 km) contraction afterward. The total volume of the dikes suggests an increase in the lunar radius of between 0.6 and 4.8 km during the expansion phase.

8.7 Dynamical History of the Moon

The orbit of the Moon is inclined slightly (5.09°) to the ecliptic, not the equator, suggestive more of a binary companion to, rather than a satellite of, the Earth. Its orbit has a perceptible eccentricity (0.0549), sufficiently large that the angular size of the Moon can be seen to vary by ~11 % from perigee to apogee. Thus, although the mean center-to-center distance from Earth is 384,400 km, the Moon's distance varies between 363,000 and 406,000 km. Moreover, because of perturbations, the longitude of the nodes regresses with a period of ~18.6 years, and the perigee advances with a period of ~9 years. The nodal regression causes a large variation in the extreme declinations of the Moon and thus of the extreme northern and southern azimuths of moonrise, especially at high latitude observing sites. See Kelley and Milone (2011), for a full discussion of this effect and its possible observation in Neolithic times and in antiquity.

The semimajor axis also varies. It appears to be increasing secularly, at a rate of ~4 cm/y (see Stephenson 1997 for a lengthy treatment of this topic), most likely due to tidal friction on Earth and the acceleration of the Moon as a consequence of a tidal bulge on Earth preceding the sublunar longitude (Sect. 6.1.1.2). The tidal forces are proportional to the inverse cube of the distance; consequently, the prediction is that the recession of the Moon will continue until the solar tides dominate.

Extrapolating into the past, the Moon must have been much nearer to the Earth, when the Earth had a shorter period of rotation. The evidence of stromatolite growth patterns has suggested a much shorter rotation period for Earth and a shorter month as recently as the Devonian.

Various scenarios for the Moon's origins have been put forward since Darwin's (1880) hypothesis that the Moon spun off from the molten Earth when the Earth formed with an angular momentum slightly above the critical value for stability at the equator.

Any model for the formation of the Moon must satisfy several constraints, including:

1. The large angular momentum of the Earth-Moon system. If the Moon did not exist, the Earth would need to rotate once every ~4^{h} to have the same angular momentum as the present Earth-Moon system. This is very different from the other inner planets.

2. The Moon's uncompressed bulk density (3,300 kg m^{-3}) which, as we have seen, is not much greater than that of the lunar crust and near that of the upper mantle, is considerably lower than that of the Earth (4,100 kg m^{-3}), and in fact is similar to the bulk density of the Earth's mantle. The difference is due primarily to the relative amounts of the most abundant heavy element, iron: the maximum Fe content of the bulk Moon is ~10 % (Canup 2004), compared to ~30 % for the Earth. The Moon is thus severely depleted in iron compared to the Earth.

3. The oxygen and silicon isotope ratios of the Moon are indistinguishable from those of the Earth, as are those of several other elements such as tungsten and titanium. These ratios differ significantly among different planets and asteroids, so the equality of these isotope ratios between the Earth and the Moon makes it unlikely that the Moon formed elsewhere in the solar system and was later captured by the Earth (aside from any dynamical problems with this model).

4. The differentiation of the Moon into a mantle and core has been dated to a time >50 My after solar system formation began (Touboul et al. 2007; Bourdon et al. 2008). (It is convenient to define "the beginning of solar system formation" as the time when the first solids, the calcium-aluminum inclusions or CAIs found in many meteorites, condensed from the solar nebula. This event has been dated to 4568.3 $\pm$ 0. 7 Ma b.p. (Burkhardt et al. 2008).) Hafnium and tungsten (Hf-W) element and isotope ratios (see below) indicate that Mars was fully-formed and differentiated into a mantle and core within ~4 My after CAI formation (Dauphas and Pourmand 2011; see also Sect. 9.3.4). Asteroids and Moon-sized planetesimals are smaller than Mars and would have formed and differentiated in <4 My, which suggests that the beginning of lunar accretion also would not have predated its completion by more than ~4 My. Thus, Moon formation was a late event in the planetary accretion process.

The time when the Moon differentiated can be found from Hf and W element and isotope ratios. ^{182}Hf is a short-lived radioactive isotope that decays into ^{182}W with a half-life of 8.9 My, whereas ^{184}W is not a decay product and so remains constant with time. The abundance ratio ^{182}W/^{184}W therefore increases with time until ^{182}Hf is extinct, while the total elemental abundance ratio Hf/W decreases with time over this same period. The latter, however, is also affected by core separation because Hf is lithophile and remains in the mantle, whereas W is siderophile and some fraction of it follows the iron into the core. Hf/W combined with other constraints can thus be used to date core formation as long as ^{182}Hf was still decaying when this occurred; or if ^{182}Hf was extinct when the core formed, then at least this fact can be established. The latter is the case for the Moon, leading to core separation >50 My after CAI formation.

Another constraint can be found from the ^{182}W/^{184}W ratio. In addition to the magma ocean, tungsten fractionation also occurred in individual magma reservoirs in the mantle; so Hf/W ratios can be different in lunar rocks derived from different melts. Nevertheless, Touboul et al. (2007) found identical ^{182}W/^{184}W ratios (equal to that of the Earth!) in 11 lunar samples having different magma source regions (two KREEP-rich samples, four low-Ti mare basalts and five high-Ti mare basalts) despite having different Hf/W ratios. The uniform ^{182}W/^{184}W ratio indicates that the lunar magma ocean remained equilibrated (and therefore convective; Sect. 8.6) until ^{182}Hf was effectively extinct. In particular, Touboul et al. (2007) find that the lunar magma ocean reached 60 % crystallization 62^{+90}_{-10} My after CAI formation. ^{147}Sm–^{143}Nd chronometry is also consistent with these ages.

5. The existence of the lunar magma ocean is attested to by the ferroan anorthositic crust (Sect. 8.6). At a solar system age of >50 My, the short-lived radioactive isotopes ^{26}Al (half-life 0.7 My) and ^{60}Fe (half-life 2.6 My) were essentially extinct, so the lunar magma ocean was not heated by their decay, and its creation required very energetic events.

6. At least some reservoirs within the Moon appear to have contents of water and other volatiles similar to those of the Earth. Results from lunar meteorites, samples returned by the Apollo and Luna programs, and remote sensing have generally been interpreted as showing that the Moon is severely depleted in volatiles. However, this deficiency may have resulted from degassing of magmas after exposure to space at the lunar surface. Recent studies of samples from which volatiles are less likely to have escaped indicate that the magma source regions had volatile contents similar to terrestrial mid-ocean ridge basalts (Hauie et al. 2011), and the lunar magma ocean may have had a water content as high as 320 ppm, about 1/3 of that of the Earth's primitive mantle (Hui et al. 2013).

Darwin's hypothesis, modified by Ringwood (1966), accounts to some extent for the similarities of composition between the Earth and the Moon, but runs into problems with angular momentum as noted above. Two other hypotheses, namely that the Earth captured the Moon (e.g., Öpik 1972) and that the Earth and Moon

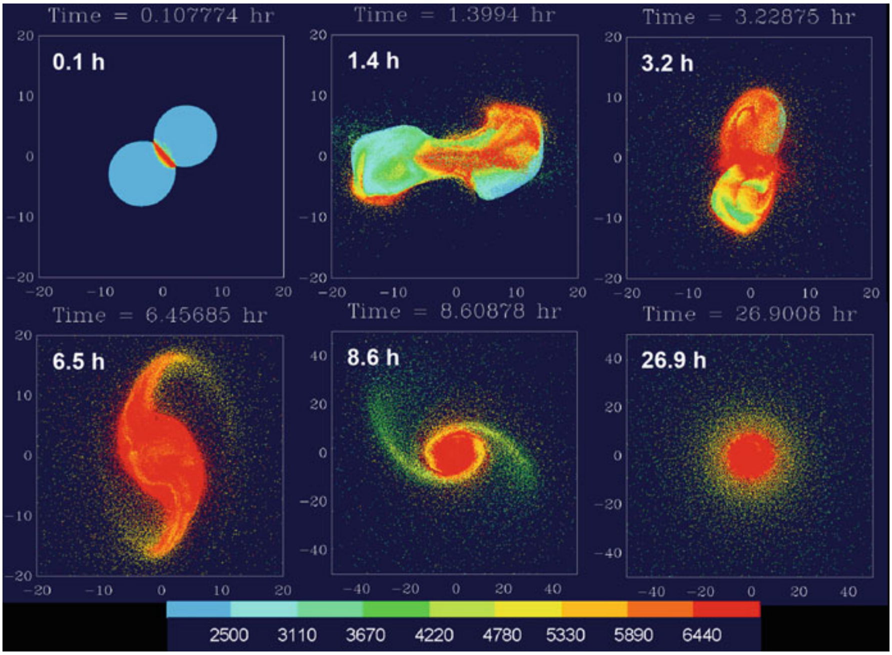


Fig. 8.13 Stages in the coalescence of the Moon after a major impact involving a proto-Earth in the early solar system. Much of the material falls back to Earth, a small amount escapes the system, and some goes into orbit around the Earth to form a disk from which the Moon later coalesces. The color bar at the bottom gives temperature from <2,500 K (*blue*) to >6,440 K (*red*). From a simulation by Robin Canup (see Canup 2012 for a full discussion) and kindly made available by her for our use here; it appears also with the permission of the American Association for the Advancement of Science

somehow were formed near each other as a true binary planet (Ruskol 1960) similarly have failed to satisfy fully both dynamical and compositional constraints. A nice discussion of the difficulties with these three hypotheses and discussion of the alternative, major impact theory described below can be found in Stevenson (1987).

The strongest and most widely accepted formation hypothesis currently involves a massive impact by a planetary-sized body on an early Earth. A computational model of such an event is illustrated in Fig. 8.13, where Robin Canup has simulated the collision over time intervals following the catastrophic collision.

Until recently, most models have assumed a Mars-sized impactor (mass ~0.1–0.2 M_T, where M_T is the total mass of the proto-Earth and impactor) striking the proto-Earth in a low-velocity, low-angle collision. The impactor's core sinks to merge with the Earth's core while its mantle only partially mixes with Earth (Cameron 1985; Benz et al. 1986; Canup 2004). Most of the rest forms a disk around the Earth from which the Moon later coalesces. This scenario helps to explain the relatively low density of the Moon (constraint 2, above), and its small iron core, while leaving the Earth-Moon system with an angular momentum similar to its present-day value.

There are two competing difficulties with the collision model as described above: Models that leave the Earth-Moon system with the required angular momentum also produce a disk (and therefore Moon) made mostly of impactor material. This leaves the Earth and Moon with different isotope ratios, violating constraint 3, above, unless the impactor has a composition very similar to the Earth—a possible but unlikely circumstance (see Milone and Wilson 2014, Ch. 15.6). Conversely, models that give the disk and Earth similar compositions leave the Earth-Moon system with excess angular momentum.

Two models were announced in 2012 that reconcile the composition and angular momentum problems. Numerical simulations by Canup (2012) indicate that, even with isotope ratios as different from the Earth as those of Mars, an impactor of mass ~0.40–0.45 M_T (i.e., ~70–80 % of the mass of the target) will result in sufficient mixing both in the disk (and therefore the resulting Moon) and in the Earth's mantle to satisfy isotope constraints. Ćuk and Stewart (2012a) look at low-mass impactors (≤ 0.1 M_T) in approximately head-on collisions with a rapidly-spinning Earth (2.3–2.9^h rotation period), dissipating about twice the kinetic energy of the older models. In one computed model, the result of an impact by a projectile of 5 % of the present Earth's mass was a disk made up of 92 % Earth material and 8 % impactor material and a final Earth's mantle made up of 98 % original Earth material and 2 % impactor material. These compositions are similar enough to satisfy the isotope constraints.

Both of these collision scenarios leave the Earth-Moon system with considerably more angular momentum than at present. It is a law of physics that the angular momentum of an *isolated system* is conserved (by definition, an isolated system is one that does not interact with any external object). For example, tidal friction within an isolated system converts mechanical energy into heat but cannot change the angular momentum of the system. Therefore, if the Earth-Moon system must reduce its angular momentum from this higher, early value to its present, lower value then it cannot be an isolated system; i.e., it must interact with an external object. The external object with the greatest influence is the Sun, but normal solar tides (i.e., excluding resonances) can reduce the angular momentum by only ~1 % over the age of the solar system (Canup 2004). However, shortly after the Moon formed, as tidal effects within the Earth-Moon system caused the Moon to migrate away from the Earth, it passed through a resonance in which angular momentum was transferred from the Earth-Moon system to the Sun much more rapidly. This was the *evection*[3] *resonance*, which occurs when the precession period of the Moon's line of apsides (the line joining the perigee and apogee points of the Moon's orbit) equals the Earth's orbital period around the Sun. In the stable configuration of the resonance, the line of apsides remains oriented so that the

[3] The present phenomenon of evection, from Latin *e* + *vehere* (*to carry away*), was discovered by Ptolemy/Hipparchus and involves the effect of the perturbations in e and ω at the Moon's perigee and apogee caused by the Sun's position at those instants. The effect is to strongly perturb the Moon's celestial longitude, λ_m. The amplitude of perturbation on λ_m at present is 1° 16′ and its period is $31\overset{d}{.}80747$.

Moon is always closest to the Sun at apogee (Yokoyama et al. 2008); the line of apsides can also librate about this configuration. The mutual perturbations between the Moon's orbital motion and the Sun's rotation thus always add, so that angular momentum transferred from the Earth to the Moon is in turn transferred from the Moon to the Sun. In the model of Ćuk and Stewart (2012a), the Moon entered the evection resonance about 9,000 years after coalescing from the disk material and exited at about 68,000 years with an angular momentum similar to that of the present Earth-Moon system. Other resonances can also transfer angular momentum out of the Earth-Moon system (Ćuk and Stewart 2012b).

Resonances have been an important factor in the orbital evolution of many objects in the solar system, not just for the Earth and Moon. According to the *Nice Model*[4] (Tsiganis et al. 2005), Uranus and Neptune, for example, formed much closer to Jupiter and Saturn than they are now, with Neptune closer to the Sun than Uranus. Slow evolution of the orbits of these four planets brought Saturn into a 1:2 orbital resonance with Jupiter about 500 My after the formation of the solar system, and the repetitive gravitational tugs on Saturn by Jupiter caused Saturn to move outward. Gravitational tugs by Saturn on Uranus and Neptune then made the orbits of the latter two much more elliptical, sending them into the population of planetesimals in the outer solar system. Gravitational interactions with the planetesimals gradually circularized the orbits of Uranus and Neptune where we find them now, with Neptune orbiting beyond the orbit of Uranus, while the planetesimals were scattered, some outward, others inward. Those that migrated inward were ejected from the solar system, were captured (Morbidelli et al. 2005), or collided with the Moon and planets (Gomes et al. 2005)—the *lunar cataclysm* or the "late heavy bombardment."

Returning to the origin of the Moon once more, we note that a challenge to the collision scenario is now posed by the results of Hauie et al. (2011) and Hui et al. (2013), which suggest very similar volatile contents in the Earth and Moon (constraint 6, above). (These results require that water found in the interiors of crystal grains or grains encased in volcanic glasses is native lunar water and not from comets, the solar wind, or terrestrial sources.) Degassing by the hot ejecta, in contrast, would be expected to produce a severely volatile-depleted Moon. Possibly the collision produced a hot, dense, volatile-rich atmosphere that enclosed both the Earth and the disk, with the volatiles being incorporated in the Moon as it coalesced, or there may have been solid fragments that retained their volatiles through disk formation and coalescence of the Moon (Hauie et al. (2011)), but specific models have yet to be developed.

This concludes our discussion of origins until Chap. 15 of Milone and Wilson (2014), when we consider the clues provided by meteorites and the small bodies of the solar system.

[4] Named for the city of Nice, in France, where principal proponents of the theory work and live.

Challenges

[8.1] Calculate the lunar crater diameter expected from the impact of an object with a 100 m diameter and density of 3,400 kg/m^3. Suppose that the object is overtaken by the Moon with a net orbital speed difference of 5 km/s.

[8.2] The crater frequency is higher over the highlands than over the lowlands of the Moon. Why? Discuss the situation in the light of the time-line of selenological history.

[8.3] The moment of inertia of the Moon is 0.391; what does this imply about the structure of the Moon?

[8.4] The angular momentum of the Earth-Moon system is 3.41×10^{34} kgm^2/s. Assuming no loss of angular momentum to the Earth-Moon system over time, find the rotation speed of the two bodies when they were in contact, under the *fission theory* for the Moon’s origin. What can you conclude assuming the *impact theory* for the Moon’s origin?

[8.5] If the acceleration of gravity is matched by the centrifugal acceleration in a critically rotating contact Earth-Moon binary, compute the breakup speed. Is the speed computed in [8.4] sufficient for fission to have occurred?

References

Andrews-Hanna, J.C., Asmar, S.W., Head III, J.W., Kiefer, W.S., Konopliv, A.S., Lemoine, F.G., Matsuyama, I., Mazarico, E., McGovern, P.J., Melosh, H.J., Neumann, G.A., Nimmo, F., Phillips, R.J., Smith, D.E., Solomon, S.C., Taylor, G.J., Wieczorek, M.A., Williams, J.G., Zuber, M.T. 2013: “Ancient Igneous Intrusions and Early Expansion of the Moon Revealed by GRAIL Gravity Gradiometry”. *Science* **339**, 675–687

Benz, W., Slattery, W.L., Cameron, A.G.W. 1986: “The Origin of the Moon and the Single-Impact Hypothesis, I”. *Icarus* **66**, 515–535

Bourdon, B., Touboul, M., Caro, G., Kleine, T. 2008: “Early differentiation of the Earth and the Moon”. *Phil. Trans. R. Soc.* **366**, 4105–4128. doi:10.1098/rsta.2008.0125

Burkhardt, C., Kleine, T., Bourdon, B., Palme, H., Zipfel, J., Friedrich, J.M., Ebel, D.S. 2008: “Hf–W Mineral Isochron for Ca,Al-rich Inclusions: Age of the Solar System and the Timing of Core Formation in Planetesimals”. *Geochim. Cosmochim. Acta.* **72**, 6177–6197

Cameron, A.G. 1985: “Formation of the Prelunar Accretion Disk”. *Icarus* **62**, 319–327

Canup, R. (2004): Simulations of a late lunar-forming impact. *Icarus* **168**, 433–456

Canup, R. (2012): Forming a moon with an earth-like composition via a giant impact. Science **338**, 1052–1054

Chenet, H., Lognonné, P., Wieczorek, M., Mizutani, H. (2006): Lateral variations of lunar crustal thickness from the Apollo seismic data set. *Earth Planet. Sci. Lett.* **243**, 1–14

Colaprete, A., Schultz, P., Heldmann, J., Wooden, D., Shirley, M., Ennico, K., Hermalyn, B., Marshall, W., Ricco, A., Elphic, R.C., Goldstein, D., Summy, D., Bart, G.D., Asphaug, E., Korycansky, D., Landis, D., Sollitt, L. (2010): Detection of water in the LCROSS ejecta plume. *Science* **330**, 463–468

Crotts, A.P.S. (2008): Lunar outgassing, transient phenomena, and the return to the moon. I. Existing data. *Astrophys. J* **687**, 692–705

Ćuk, M., Stewart, S.T. (2012): Making the moon from a fast-spinning earth: a giant impact followed by resonant despinning. *Science* **338**, 1047–1052

Ćuk, M., Stewart, S.T. Resonances and the Angular Momentum of the Earth-Moon System. *Early Solar System Impact Bombardment II*, Feb. 1-3, Houston, abstract 4006 (2012b)

Darwin, G.H.: On the secular change in elements of the orbit of a satellite revolving around a tidally distorted planet. *Phil. Trans. R. Soc. Lond.* **171**, 713–891 (1880)

Dauphas, N., Pourmand, A.: Hf–W–Th evidence for rapid growth of mars and its status as a planetary embryo. *Nature* **473**, 489–493 (2011)

Elardo, S.M., Draper, D.S., Shearer Jr., C.K.: Lunar magma ocean crystallization revisited: bulk composition, early cumulate mineralogy, and the source regions of the highlands Mg-suite. *Geochim. Cosmochim. Acta* **75**, 3024–3045 (2011)

Gagnepain-Beyneix, J., Lognonné, P., Chenet, H., Lombardic, D., Spohnd, T.: A seismic model of the lunar mantle and constraints on temperature and mineralogy. *Physics of the Earth and Planetary Interiors* **159**, 140–166 (2006)

Garcia, R.F., Gagnepain-Beyneix, J., Chevrot, S., Lognonné, P.: Very preliminary reference Moon model. *Physics of the Earth and Planetary Interiors* **188**, 96–113 (2011). Note: Erratum, *ibid.*, 2012

Garrick-Bethell, I., Fernandes, V.A., Weiss, B.P., Shuster, D.L., Becker, T.A. 4.2 Billion Year Old Ages from Apollo 16, 17, and the Lunar Farside: Age of the South Pole-Aitken Basin? in Workshop on the Early Solar System Impact Bombardmen*t,* p. 34–35. *LPI Contribution* No. 1439, Lunar and Planetary Institute, Houston (2008)

Gomes, R., Levison, K., Tsiganis, K., Mobidelli, A.: Origin of the late heavy bombardment period of the terrestrial planets. *Nature* **435**, 466–469 (2005)

Hiesinger, H., van der Bogert, C.H., Pasckert, J.H., Schmedemann, N., Robinson, M.S., Jolliff, B., and Petro, N. New Crater Size-Frequency Distribution Measurements of the South Pole-Aitken Basin. *43rd Lunar and Planetary Science Conference*, held March 19–23, 2012 at The Woodlands, Texas. LPI Contribution No. 1659, id.2863 (2012)

Hiesinger, H., Head III, J.W., Wolf, U., Jaumann, R., Neukum, G.: Ages and stratigraphy of mare basalts in Oceanus procellarum, mare nubium, mare cognitum, and mare insularum. *J. Geophys. Res.* **108**(E7), 5065 (2003). doi:10.1029/2002JE001985 (27 pp.)

Hood, L.L., Mitchell, D.L., Lin, R.P., Acuna, M.H., Binder, A.B..: Initial measurements of the lunar induced magnetic dipole moment using lunar prospector magnetometer data. *Geophys. Res. Lett.* **26**, 2327–2330 (1999)

Hauie, E.H., Weinreich, T., Saal, A.E., Rutherford, M.C., Van Orman, J.A.: High Pre-eruptive water contents preserved in lunar melt inclusions. *Science* **333**, 213–215 (2011)

Hui, H., Peslier, A.H., Zhang, Y., Neal, C.R.: Water in lunar anorthosites and evidence for a wet early moon. *Nature Geoscience* (2013). doi:10.1038/ngeo1735 (4 pages)

Jolliff, B.L., Gillis, J.J., Haskin, L., Korotev, R.L., Wieczorek, M.A.: Major lunar crustal terranes: surface expressions and crust-mantle origins. J. Geophys. Res. **105**, 4197–4216 (2000)

Kelley, D.H., Milone, E.F.: *Exploring Ancient Skies* Springer, New York (2011)

Korotev, R.L.: The great lunar hotspot and the composition and origin of the Apollo mafic ("LKFM") impact-melt breccias. J. Geophys. Res. **105**, 4317–4345 (2000)

Korotev, R.L., Gillis, J.J.: A New look at the Apollo 11 regolith and KREEP. *J. Geophys. Res.* **106**, 12,339–12,353 (2001)

Liu, Y., Guan, Y., Zhang, Y., Rossman, G.R., Eiler, J.M., Taylor, L.A.: Direct measurement of hydroxyl in the lunar regolith and the origin of lunar surface water. Nature Geoscience **5**, 779–781 (2012)

Milone, E.F., Wilson, W.J.F. 2014: *Solar System Sstrophysics: Planetary Atmospheres and the Outer Solar System.* Springer, New York

Morbidelli, A., Levison, H.K., Tsiganis, K., Gomes, R.: Chaotic capture of Jupiter's Trojan asteroids in the early solar system. *Nature* **435**, 462–465 (2005)

Nakamura, Y.: Farside deep moonquakes and deep interior of the moon. *J. Geophys. Res.* **110**, E01001 (2005). doi:10.1029/2004JE002332 (12 pp.)

Öpik, E.J.: Comments of lunar origin. *Irish Astr. J.* **10**, 190–238 (1972)

Ringwood, A.E.: Chemical evolution of the terrestrial planets. *Geochima. Cosmochim. Acta.* **30**, 41–104 (1966)

Ruskol, E.L.: Origin of the moon. I. *Soviet. Astronomy, AJ* **4**, 657–668 (1960)

Stephenson, F.R.: *Historical eclipses and Earth's rotation.* University Press, Cambridge (1997)

Stevenson, D.J.: Origin of the moon—the collision hypothesis. *Annu. Rev. Earth Planet. Sci.* **15**, 271–315 (1987)

Stoffler, D., Knoll, H.D., Marvin, U.B., Simonds, C.H., and Warren, P.H. Recommended Classification and Nomenclature of Lunar Highland Rocks: a Committee Report. in *Proceedings of the Conference on the Lunar Highland Crust* (Pergamon, New York), 51–70 (1980)

Sunshine, J.M., Farnham, T.L., Feaga, L.M., Groussin, O., Merlin, F., Milliken, R.E., A'Hearn, M.F.: Temporal and spatial variability of lunar hydration as observed by the deep impact spacecraft. *Science* **326**, 565–568 (2009)

Touboul, M., Kleine, T., Bourdon, B., Palme, H., Wieler, R.: Late formation and prolonged differentiation of the moon inferred from W isotopes in lunar metals. Nature **450**, 1206–1209 (2007)

Tsiganis, K., Gomes, R., Morbidelli, A., Levison, H.F.: Origin of the orbital architecture of the giant planets of the solar system. *Nature* **435**, 459–461 (2005)

Warren, P.H.: The moon. *Treatise on Geochemistry* **1**, 559–599 (2003)

Weber, R.C., Lin, P.-Y., Garnero, E.J., Williams, Q., Lognonné, P.: Seismic detection of the lunar core. *Science* **331**, 309–312 (2011)

Wieczorek, M.A., Jolliff, B.L., Khan, A., Pritchard, M.E., Weiss, B.P., Williams, J.G., Hood, L.L., Righter, K., Neal, C.R., Shearer, C.K., McCallum, I.S., Tompkins, S., Ray Hawke, B., Peterson, C., Gillis, J.J., Bussey, B.: The constitution and structure of the lunar interior. *Reviews in Mineralogy & Geochemistry*. **60**, 221–364 (2006)

Wieczorek, M.A., Neumann, G.A., Nimmo, F., Kiefer, W.S., Taylor, G.J., Melosh, H.J., Phillips, R.J., Solomon, S.C., Andrews-Hanna, J.C., Asmar, S.W., Konopliv, A.S., Lemoine, F.G., Smith, D.E., Watkins, M.M., Williams, J.G., Zuber, M.T.: The crust of the moon as seen by GRAIL. *Science* **339**, 671–675 (2013)

Wilcox, B.B., Robinson, M.S., Thomas, P.C., Hawke, B.R.: Constraints on the depth and variability of the lunar regolith. Meteoritics & Planetary Science **40**, 695–710 (2005)

Williams, J.G., et al.: Lunar rotational dissipation in solid body and molten core. *J. Geophys. Res.* **106**, 27,933–27,968 (2001)

Wood, J.A., Lunar Sample Analysis Planning Team: Third lunar science conference: primal igneous activity in the outer layers of the Moon generated a Feldspathic crust 40 kilometers thick. *Science* **176**, 975–981 (1972). No. 4038

Yokoyama, T., Vieira Neto, E., Cabo Winter, O., Merguizo Sanchez, D., de Oliveira Brasil, P.I.. On the Evection Resonance and Its Connection to the Stability of Outer Satellites. in *Mathematical Problems in Engineering* Hindawi Publishing Corporation (16 pages) (2008)

Chapter 9
Surface Science of the Terrestrial Planets

In this chapter we explore the surfaces as well as the interiors of Mercury, Venus, and Mars, and compare their properties to those of the Earth and Moon, which we have already examined. In Milone and Wilson (2014, Chap. 10), we study the nature of atmospheres and ionospheres with tools of physics and chemistry; in Chap. 11 we consider the magnetospheres.

The orbital, physical, and photometric properties of the terrestrial planets are summarized in Table 9.1. In the order given in column 1, these are: the semi-major axis, a; the eccentricity, *e*; the inclination with respect to the ecliptic, *i*; the sidereal period in units of the Earth's sidereal year; the mass in Earth masses (Earth's in kg); the radius in Earth radii (Earth's in km); the sidereal rotation period in mean solar days; the oblateness, departure from a spherical shape; the Legendre polynomial coefficients J (see Sects. 5.3, 5.4.3, 5.6); the mean density; the mean surface gravity at the planet's equator, g_{eq}; the escape velocity from the surface, v_∞; the calculated V magnitude as if the object were located 1 au from the Sun, and if seen at opposition from a distance of exactly 1 au; the (B-V) and (U-B) color indices (larger indicates redder); the visual (V magnitude) geometric albedo (the ratio of the V light reflected from the apparent disk of the planet compared to that of a perfectly reflecting Lambert disk; the Bond albedo, the ratio of the total light reflected to that incident on the planet; the amount of radiant energy falling on a square meter of surface; the (predicted) equilibrium temperature; and an observed temperature, in kelvin. The photometric data are from the 2012 Astronomical Almanac; the Bond albedo for Venus is from Moroz et al. (1985), while those for Mercury, the Earth and Mars are from NASA's planetary factsheets: http://nssdc.gsfc.nasa.gov/planetary/planetfact.html.

We now consider the planets in their heliocentric order, starting with Mercury.

E.F. Milone and W.J.F. Wilson, *Solar System Astrophysics: Background Science and the Inner Solar System*, Astronomy and Astrophysics Library,
DOI 10.1007/978-1-4614-8848-4_9, © Springer Science+Business Media New York 2014

Table 9.1 Bulk properties of the terrestrial planets

Properties	Mercury	Venus	Earth	Mars
Orbital data				
a (au)	0.3871	0.7233	1.0000	1.5236
e	0.2056	0.0068	0.0167	0.0933
i (to ecliptic)	$7\overset{\circ}{.}0042$	$3\overset{\circ}{.}3947$	$0\overset{\circ}{.}0017$	$1\overset{\circ}{.}8487$
Sidereal period	$0\overset{y}{.}2408$	$0\overset{y}{.}6152$	$1\overset{y}{.}0000 = 365\overset{d}{.}256363$	$1\overset{y}{.}8807$
Physical data:				
Mass ($M_{\oplus}$)	0.0553	0.8150	$1 = 5.972 \times 10^{24}$ kg	0.1074
R_{eq} ($R_{\oplus}$)	0.3825	0.9488	1 = 6378.137 km	0.5325
P_{rotation}[a]	58.6462	−243.023[b]	0.99726963	1.02595676
Oblateness[c]	0	0	0.0035364	0.006772[d]
J_2	. . .	2.70×10^{-5}	1.083×10^{-3}	1.964×10^{-3}
J_3	. . .	. . .	-2.64×10^{-6}	3.6×10^{-5}
J_4	. . .	. . .	-1.61×10^{-6}	. . .
$<\rho>$ (kg/m^3)	5,430	5,240	5,515	3,940
$<g_{\text{eq}}>$ (m/s^2)	3.703	8.871	9.798	3.711
v_{∞} (km/s)	4.25	10.36	11.18	5.02
Photometric data:				
V (1,0)	−0.42	−4.40	−3.86	−1.52 [−2.01][e]
(B-V)	+0.93	+0.82	. . .	+1.36
(U-B)	+0.41	+0.50	. . .	+0.58
Radiative energy data:				
Albedo (geom., V)	0.138	0.67	0.367	0.15
Albedo (Bond, bol.)	0.119	0.76	0.306	0.25
Incident solar Flux (W/m^2)	9,076	2,600	1,360	586
T_{eq} (predicted effective temperature)	515 K[f]	272 K[f]	254 K[g]	210 K[g]
T_{eff} (observed)	700 K[h]	740 K	288 K	210 K

[a] Sidereal, in mean solar days of length 86,400 s

[b] Mean rotation period over 16 years, from *Magellan* (1990–1992) and *Venus Express* (2006–2008) data

[c] $\varepsilon = (R_{\text{eq}} - R_{\text{pol}})/R_{\text{eq}}$

[d] N pole radius result; S pole result: 0.005000

[e] [V_0, The photometric V magnitude at mean opposition]. V(1,0) is the V magnitude that the planet would have if it were viewable at 1 au from the Sun and 1 au from the Earth, at a phase angle of 0 (see Sect. 5.5.2). The astronomical unit (au) is formally defined as 149597870.7 km

[f] "Slow rotator" case

[g] "Fast rotator" case

[h] Subsolar temperature at perihelion

9.1 Mercury

9.1.1 Visibility

Mercury is the smallest planet in the solar system,[1] with a radius of 2,439 km, or 0.38 $\mathfrak{R}_{\oplus}$. It was known in antiquity as the messenger of the gods, because of its apparent rapid shuttle motion back and forth around the Sun, overtaking planet after planet during the course of its elaborate motions. In ancient Greece it was known as Hermes when seen in the evening sky, and Apollo when seen before sunrise. In the ancient Egyptian culture the two apparitions were sometimes given the names Horus and Seth, and among the Hindus, Raulineya and Buddha.

Its angular diameter varies between about 5 and 11 arc-secs as viewed from Earth. At maximum elongation, a typical brightness is ~0.5 magnitude and a diameter of ~8 arc-secs.

Mercury is difficult to see because of its proximity to the Sun. Its maximum elongation is ~28° (see Fig. 9.1). Abell (1969, p. 296) commented that despite the fact that it is the seventh brightest object (after the Sun, Moon, Venus, Mars, Jupiter, and Sirius) in our sky, ". . . most people—including even Copernicus, it is said—have never seen Mercury."

Its computed magnitude at superior conjunction (when it is not readily visible) is -2.2, which potentially makes it even brighter than Sirius. Its proximity to the Sun means that it must be observed in daylight, or twilight, and, if it is to be observed when the Sun is below the horizon, it is seen through high air mass. Not

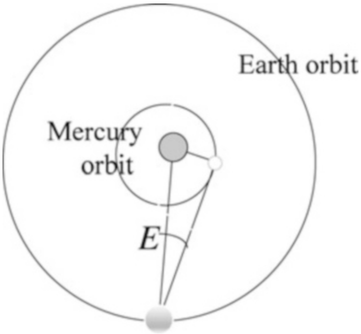


Fig. 9.1 Mercury, at a "maximum western" *elongation* (E). As the Earth rotates CCW, Mercury in this configuration rises before the Sun, hence west of the Sun in our sky. Note that this particular maximum elongation is not optimal for observing Mercury

[1] Pluto, which is smaller than Mercury and the Earth's moon, and may not be the largest icy body beyond the orbit of Neptune in the outer solar system, was reclassified as a "dwarf planet" by the International Astronomical Union in 2006 (see Milone and Wilson 2014, Chaps. 13–16).

surprisingly, the best ground-based views have not been especially good. Consequently, although Mercury has been imaged from the ground on occasion, the best views we have of Mercury are from spacecraft.

9.1.2 Mercury's Orbit

Mercury has a semi-major axis of 0.38710 au and the inclination of its orbit to the ecliptic ($7\overset{\circ}{.}005$) is larger than that of all other planets (dwarf planets excluded). Mercury's mean orbital velocity is 47.89 km/s. It is interesting and suggestive that the inclination is similar to the angle between the Sun's equatorial plane and the ecliptic, $7\overset{\circ}{.}25$.

Mercury also has the highest eccentricity of all the (major) planets, $e = 0.2056$. For the epoch JDN 2456310.5 = January 19, 2013, 0^{h} U.T., other elements had the values: the longitude of its ascending node, $\Omega = 48\overset{\circ}{.}3147$ and the longitude of perihelion, $\varpi = 77\overset{\circ}{.}4783$. As with the other planets, all of its elements vary with time. Mercury's *a* and *i*, given above, are accurate for this epoch. The closeness to the Sun produces an additional source of perturbation to Mercury's osculating orbits.

Due to its high eccentricity, Mercury's orbit has been a testbed for gravitational theories. Misner et al. (1973, p. 1113) indicate that the line of apsides rotates forward (in the direction of orbital motion) at a rate of 55.9974(41) arc-sec/y (the parenthesis contains the uncertainty in units of the last decimal place: ±0.0041). Of this amount, 50.25645(50) arc-sec/y is due to "general precession" of the vernal equinox, a "contribution to the shift caused by the observer not being in an inertial frame far from the Sun."

Perturbations of other planets contribute 5.3154(68) arc-sec/y. The residual excess shift is 0.4256(94) arc-sec/y, which is very close to the amount predicted by general relativity (0.42980 arc-sec/y). Within the 1-sigma (standard deviation) uncertainty, the values agree.

Leverrier, in the nineteenth century, attributed this excess to the perturbations of a planet interior to Mercury, which he named *Vulcan*. This predicted planet has since been proven fictitious; and although there are asteroids in the inner portion of the solar system, their small masses and complex orbital movements make any of them unlikely to be responsible for the effect.

As noted in Sects. 3.8 and 5.3, this type of perturbation can, in principle, arise from an aspherical mass distribution in the primary, in this case the Sun, due to rotation (i.e., rotational flattening or oblateness); see Dicke and Goldenberg (1967). Expressed as the difference between the equatorial and polar angular radii of the Sun, the expected solar oblateness is $\Delta r_{\mathrm{surf}} = 7.8$ milliarcseconds (mas) if the entire Sun rotates at the same angular speed as the surface (the surface rotational speed is ~2 km/s at the equator); see Sect. 4.2. An observed oblateness significantly larger than this would indicate a significant internal oblateness due to rapid internal rotation. This in turn would account for part of the residual excess shift and reduce the agreement between the observed excess shift and the shift predicted by general relativity. Fivian et al. (2008), using observations from the RHESSI satellite, measured a value of $\Delta r = r_{equatorial} - r_{polar} = 8.01 \pm 0.14$ mas after allowing for latitude-dependent magnetic effects on solar brightness, and Kuhn et al. (2012) found 7.20 ± 0.49 mas using the Solar Dynamics Observatory; both of these values are very close to the 7.8 mas value expected from the observed surface rotation. From their value for Δr and a mean solar radius $r_0 = 959.63$ arc-sec from Cox (2000), Fivian et al. (2008) obtain a value for the J_2 term in the expression for the solar potential of $J_2 = 2/3(\Delta r - \Delta r_{surf})/r_0 = (1.46 \pm 1.0) \times 10^{-7}$. The corresponding contribution to the advance in the perihelion of Mercury is ~0.0002 arc-sec/y, which is significantly smaller than the uncertainties quoted above. Thus, the contributions to Mercury's apsidal motion from other causes are insufficient to cast doubt on the value predicted by general relativity.

9.1.3 Mercury's Physical Properties

Our first detailed views of Mercury were from NASA's *Mariner 10* spacecraft, which flew past Mercury three times: March 29 and September 21, 1974, and March 16, 1975. Because of *Mariner 10*'s orbit and Mercury's 3:2 spin-orbit coupling, described below, *Mariner* saw the same side of Mercury all three times.

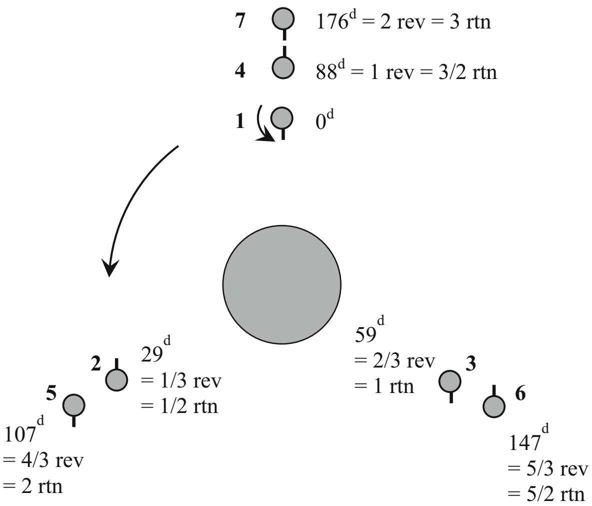
Many more detailed views have been provided NASA's *MESSENGER* (MErcury Surface, Space ENvironment, GEochemistry, and Ranging) spacecraft, which conducted three flybys of the planet (January 14 and October 6, 2008, and September 29, 2009) before being placed in an almost-polar orbit (83.5° inclination) around Mercury on March 18, 2011. The ~12-h orbit is highly elliptical to avoid being overheated by Mercury's hot surface, with an initial periapsis at an altitude of 200 km over the northern hemisphere and apoapsis at ~15,200 km over the southern hemisphere. As a result, the northern hemisphere has been studied in much more detail than the southern; e.g., the laser altimeter can be used only below an altitude of 1,500 km pointing straight down, or 1,000 km at an angle of 40° from the vertical, so it can provide topographic information only for the northern hemisphere.

9.1.3.1 Spin-Orbit Coupling

Mercury has a heavily cratered surface, and its features as viewed telescopically are only poorly resolved smudges. Between 1882 and 1890, Giovanni Schiaparelli observed the planet closely and reported a rotation period of 88^d. The proximity to the Sun was expected to result in tides on Mercury strong enough to cause a 1:1 spin–orbit lock, so this result was accepted for decades.

However, radio data from Mercury in 1962 suggested the presence of thermal radiation from the darkened disk of Mercury, an unexpected result if Mercury were indeed locked 1:1, because the "dark" side should be perpetually in darkness except for librational effects.

Fig. 9.2 The 3:2 spin–orbit coupling of Mercury and consequences for a solar day at a particular location on Mercury (indicated by the very tall flagpole on the planet)



In 1965, R. Dyce and G. Pettingill used the Arecibo Radio Telescope to send radar pulses to Mercury and receive the return echo. The time delays and Doppler shifts were analyzed, resulting in a rotation period of 59^d compared to the 88^d orbital period. Precise values of these quantities are:

$$P_{\text{rotation}} = 58\overset{d}{.}6462$$
$$P_{\text{revolution}} = 87\overset{d}{.}969$$

Thus Mercury rotates three times on its axis for every two orbital revolutions, a state known as 3:2 spin-orbit coupling. The geometry is shown in Fig. 9.2.

The reason for Schiaparelli's conclusion is seen in Fig. 9.3, based loosely on Fig. 4.9 of Consolmagno and Schaefer (1994). On average, an observer on the Earth gets a relatively good look at the same region of Mercury only every fourth revolution (at every second 88^d interval the observer sees a different region of Mercury—but it too will be the same every fourth revolution). Consequently, because the aphelion apparitions of Mercury were more favorable for observation, Schiaparelli's data were aliased in such a way as to convince him, given the lock-in model for the Earth's moon, that he had the correct explanation. His data do indeed fit the 59^d rotation interval. It is seen that one of two longitude regions always faces the Sun at perihelion: longitude 0° and 180°. The *Caloris Basin* (Fig. 9.5a), an aptly named multi-ringed impact feature, is near, although not centered on, longitude 180°.

The mean solar day length is found from the relative angular rate formula:

$$\omega_{msd} = \omega_{rtn} - \omega_{rev} \tag{9.1}$$

$$2\pi/P_{msd} = 2\pi/P_{rtn} - 2\pi/P_{rev} \tag{9.2}$$

from which we find

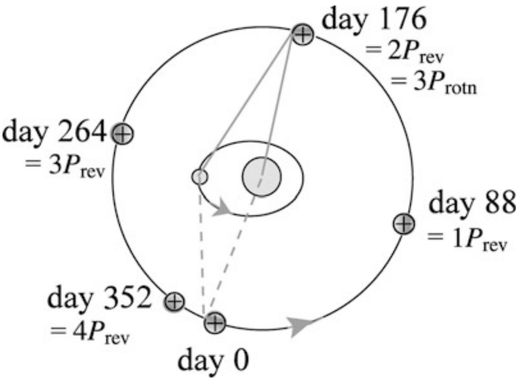


Fig. 9.3 Similar configurations of Mercury are seen from the Earth every two Mercurian sidereal years, which is equal to three Mercury rotations. Therefore during successive intervals of favorable maximum elongation (i.e., when Mercury is at maximum elongation at aphelion), the same region of Mercury faces the Sun. Based loosely on Consolmagno and Schaefer (1994, Fig. 4.9)

$$P_{\rm msd} = 179^{\rm d}$$

If Mercury were moving in a circular orbit, the Sun would be above the equatorial horizon for half of this interval, or $89^{\rm d}$. However, at perihelion Mercury's large eccentricity causes it to sweep around the Sun faster than the rotation ($\omega_{\rm rev} > \omega_{\rm rtn}$). Therefore, the Sun on Mercury would normally be seen to move from east to west, then, near perihelion, the following sequence would be observed:

Sun becomes stationary
Sun moves eastward
Sun becomes stationary
Then the Sun moves westward again, repeating the cycle

The visibility of the Sun in the sky for other latitudes on Mercury differs only slightly from that at the equator in ways which are better illustrated for planets with much larger differences between their rotational and revolution axes, so we leave this description to the interested reader! We now turn to a discussion of the physical characteristics of the planet itself. Further discussion of such topics as the surface reflectance, and the crater size/frequency distribution, compared to those of the Moon, can be found in Strom and Sprague (2003).

9.1.3.2 Size, Shape, Mass and Density

The radius of Mercury, from angular measurement via *Mariner 10*, is 2439.7 km. The slow rotation explains the planet's small amount of *flattening* or *oblateness,*

$$\left(\mathfrak{R}_{\mathrm{eq}} - \mathfrak{R}_{\mathrm{pol}}\right)/\mathfrak{R}_{\mathrm{eq}} = 0.0 \tag{9.3}$$

The equator is slightly elliptical, with the major axis 1.25 km longer than the minor axis and pointing toward a longitude ~19° W of the prime meridian (Zuber et al. 2012). The prime meridian on Mercury (longitude 0°) has been chosen to coincide with one of the hot poles; i.e., one of the two points where the Sun is directly overhead when Mercury is at perihelion (note positions 1, 4 and 7 in Fig. 9.2).

Mercury's obliquity, the angle between the plane of rotation and the orbital plane, or the angle between the rotation axis and the orbit normal, is only 2.06 ± 0.1 arc min (Smith et al. 2012).

Mercury has no moon, but its mass could be calculated from *MESSENGER*'s orbit. However, the value was already well-known from a 1968 approach of the asteroid Icarus to within 16 million kilometers of Mercury, and subsequent space probes to Venus and Mercury were subject to perturbations from Mercury. These particular events and ongoing *n*-body numerical simulations of all the planets in the solar system produce more and more refined values of the mass of Mercury (as well as of the other planets). The 2012 Astronomical Almanac gives the value:

$$3.3010 \times 10^{23}\mathrm{kg}$$

The combination of mass and volume provides the mean density:

$$<\rho> \;=\; 5430\ \mathrm{kg/m^3} = \; 5.430\ \mathrm{g/cm^3}$$

This compressed mean density is higher than that of any other planet except Earth, which suggests that there is a substantial amount of iron in the interior. Moreover, large amounts of this iron must be in metallic form, which has a closely packed crystal structure. Fayalite (Fe_2SiO_4, in the olivine group), the most iron-rich of the common silicates, has a density of only 4,200 kg/m^3, insufficient to explain Mercury's density.

Mercury has a low geometric albedo,[2] 0.138, and mean color indices: (B-V) = 0.93 and (U-B) = 0.41. These colors are significantly redder than sunlight (0.65, 0.20, respectively). The lightly-cratered plains are not, however, much darker than the heavily-cratered areas (unlike the Moon where the mare basalts are much darker than the heavily-cratered highlands; see Taylor 1992, p. 193).

[2] The geometric albedo is defined in the Astronomical Almanac as the ratio of the illumination of the planet at zero phase angle (i.e., the brightness as viewed from the light source) to that of a pure white Lambert plane surface (or Lambertian surface) of the same radius and position as the planet. A Lambert plane surface is one for which the reflected radiant intensity (or the reflected luminous intensity), I, is directly proportional to the cosine of the angle, θ, between the observer's line of sight and the surface normal: $I(\theta) = I(0) \cos \theta$.

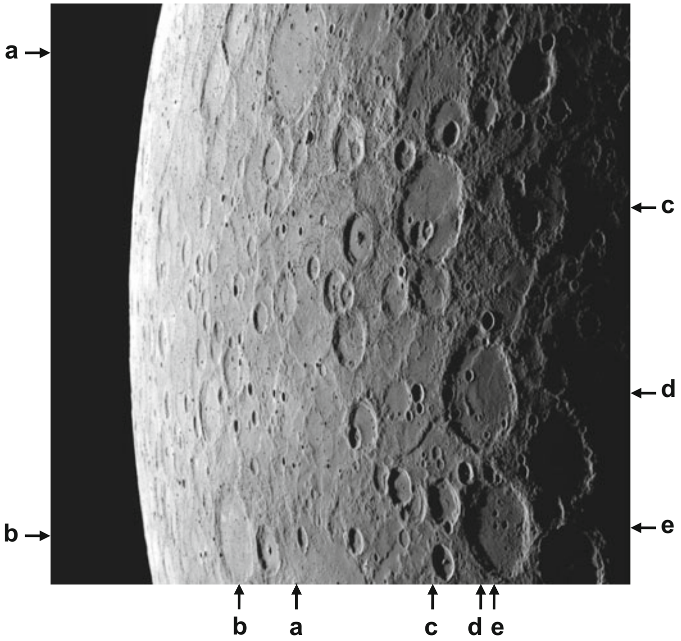


Fig. 9.4 Mercury's heavily-cratered surface, photographed by the *MESSENGER* spacecraft. North is up. The crater with the prominent central peak, just above center, is Bashō (dia. 75 km). Other, large craters in the image are, a: Ustad Isa (dia. 138 km), b: Takayoshi (dia. 136 km), c: Barma (dia. 123 km), d: Milton (dia. 181 km), e: Liang K'ai (dia. 145 km). Image PIA15373, courtesy of NASA/Johns Hopkins University Applied Physics Laboratory/Carnegie Institution of Washington

The centre of mass is offset from the centre of figure by 0.234 km toward longitude 55° in the equatorial plane (Zuber et al. 2012). Factors that could contribute to such an offset include large-scale variations in crustal thickness or density, mantle density, and topography of the core-mantle boundary.

9.1.3.3 Global Topography and Terrain

Figure 9.4 shows a large-scale view of a region in Mercury's southern hemisphere.

Results from NASA's *MESSENGER* spacecraft show that Mercury's surface consists of basalt from flood volcanism, with subsequent modification by impact craters, impact basins, and coverage by impact ejecta. No evidence has yet been found for an anorthositic primary crust such as was produced on the Moon by flotation of feldspar in a magma ocean.

Based on morphology, relative reflectance in the wavelength range 430–1,020 nm, and spectral slope in the same wavelength range (the slope of intensity of reflected light *vs.* wavelength; steeper spectral slope means redder), most of Mercury's surface can be divided into three terrain types (Denevi et al. 2009 and references therein):

1. Smooth plains:
   - Widespread and globally distributed, covering ~40 % of Mercury's surface; individual deposits can be up to 1.7 million km^2 (the Caloris basin interior plains), rivaling the largest flood basalt units on the Earth or Moon
   - The lowest density of cratering, therefore the youngest of the three terrain types
   - Typically fill low-lying areas such as impact craters and basins
   - Several generations of smooth plains are evident in many areas
   - Can be up to 5 km thick, as shown by crater excavation

   Smooth plains are divided into three subtypes:

   High-reflectance red plains (HRP):

   - The most conspicuous of the smooth plains: reflectances up to 20 % above the global mean
   - Relatively steep spectral slopes (i.e., redder)
   - Generally sharp color and morphologic boundaries
   - Strong evidence of having originated from flood volcanism

   Intermediate plains (IP):

   - Reflectance and color properties similar to the global mean
   - Sharp morphologic boundaries
   - Sharp color boundaries where they overlie LRM (see below)
   - Strong evidence of having originated from flood volcanism

   Low-reflectance blue plains (LBP):

   - Reflectance 15 % below the global mean
   - Spectral properties intermediate between those of the IP and low-reflectance material (see below)
   - Origin uncertain; e.g., the circum-Caloris plains, the largest LBP expanse, in some respects seem to be related to Caloris ejecta, so they should be the same age as the impact, yet crater counts indicate that they are younger than the Caloris interior plains which in turn are younger than the Caloris rim, suggesting later emplacement, possibly by flood volcanism

2. Intermediate terrain (IT):
   - Higher crater density (and therefore older) than the smooth plains
   - Reflectance and color properties similar to the global mean, with moderate variation
   - Generally correspond to the "heavily cratered terrain" and "intercrater plains" in *Mariner 10* images
   - The intercrater plains may be older, more degraded smooth plains formed by flood volcanism during or at the end of the late heavy bombardment; however, an origin as molten basin ejecta is not ruled out.

3. Low-reflectance material (LRM):

   - Covers at least 15 % of Mercury's surface; individual regions can cover >4 million km^2
   - Reflectance as low as 30 % below the global mean
   - ~5 % Shallower spectral slope than the HRP (therefore classified as "blue" relative to HRP), with a relative upturn at wavelengths below ~500 nm
   - No distinctive morphologic characteristics; except for its color properties, much of it would be mapped as IT
   - The LRM appears to consist of subsurface material thrown out as crater or basin ejecta, with depths of origin from several kilometers to as much as 25 km. It therefore appears to represent a component of the lower crust or upper mantle that was redistributed as a veneer on the surface. Some diffuse LRM deposits, however, may be intrusive or extrusive deposits of magma that have been degraded and mixed by impacts.
   - Many large impact craters do not show LRM in their ejecta, suggesting that the source material is not distributed uniformly throughout the crust either horizontally or vertically.

The *dynamic range* (the vertical distance from the lowest to the highest elevation) of Mercury's topography, at least in the northern hemisphere where *MESSENGER*'s laser altimeter can be used, was measured with respect to a reference radius of 2,440 km. The extremes are: 5.824 km below, and 4.024 km above, for a dynamic range of 9.848 km (Zuber et al. 2012). This range is significantly smaller than on the other three terrestrial planets or the Moon:

- Venus, 13.8 km (Diana Chasma, −2.9 km; Maxwell Montes, + 10.9 km)
- Earth, 19.84 km (Challenger Deep, −10.99 km; Mt. Everest, +8.85 km)
- Moon, 19.90 km (low point, −9.115 km; high point, +10.786 km)
- Mars, 29.43 km (Hellas Basin, −8.20 km; Olympus Mons, +21.23 km).

The dynamic range is influenced by, among other things,

- Tectonic activity, which produces mountains and chasms (Maxwell Montes and Diana Chasma on Venus, the Himalayas and the Marianas Trench on Earth). Mercury's mantle is <400 km thick (Sect. 9.1.4), which may have limited the intensity and duration of any tectonic activity.
- Shield volcanism and large-scale extension. On Mars, Olympus Mons is the largest shield volcano in the solar system, whereas shield volcanism is absent on Mercury. However, if we exclude Olympus Mons and the Tharsis Rise, then the dynamic range of the remaining topography on Mars is approximately the distance from the floor to the rim of Hellas Basin, ~9 km, which is similar to that of Mercury.
- The value of the gravitational acceleration, g_{rel}, relative to that of Earth: Moon, 0.17 $g_{\oplus}$; Mercury, 0.38 $g_{\oplus}$; Venus, 0.91 $g_{\oplus}$; Mars, 0.38 $g_{\oplus}$. The higher the value of g, the stronger the forces acting to reduce dynamic range.

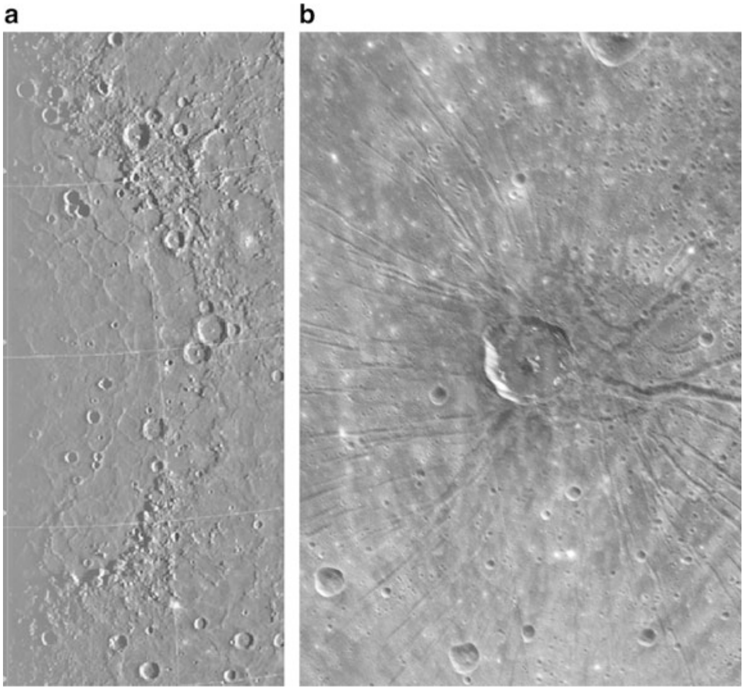


Fig. 9.5 (a) A rim of the Caloris Basin in a portion of a gridded image from a *Mariner 10* flyby of Mercury. JPL/NASA/Northwestern University image PIA024139. (b) A set of radiating troughs near the center of Caloris Basin, photographed during the first Mercury flyby of the *MESSENGER* spacecraft, January, 2008, in an area that was in darkness during the *Mariner 10* flybys. The troughs, nicknamed "the spider," appear to have formed by extension and splitting of the floor material of the basin. A 40-km diameter crater lies close to the center of the pattern, but it is not clear if the troughs resulted from the impact or formed earlier, or perhaps a combination of the two. Image courtesy of NASA/Johns Hopkins University Applied Physics Laboratory/Carnegie Institution of Washington

9.1.3.4 Topographic Features

Caloris Basin

Caloris Basin (Fig. 9.5a), 1,550 km in diameter, is the best-preserved and therefore probably the youngest large impact basin on Mercury. It is intermediate in size between Mare Imbrium and South Pole-Aitken on the Moon (1,150 km and 2,500 km in diameter, respectively). The floor of the basin is covered with high-reflectance smooth plains (HRP; see Sect. 9.1.3.3) created by flood volcanism.

Figure 9.6 plots the elevation vs. position along an approximately NW-SE transect through the centre of Caloris Basin (the transect is shown in Figure 3 of Oberst et al. 2010). A pair of rises that form an undulation of amplitude ~1.5 km and wavelength ~900 km is evident in the graph, with the floor of the basin at the more

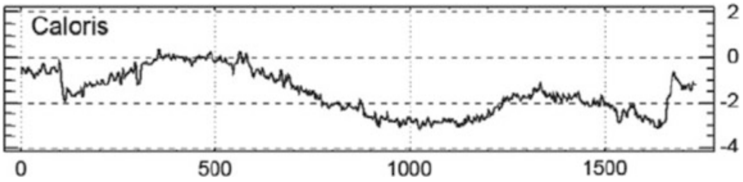


Fig. 9.6 Elevation profile in a northwest-to-southeast direction through Caloris Basin, showing elevation in km on the vertical scale *vs.* position in km on the horizontal scale (from Fig. 8 of Oberst et al. 2010, with permission of the authors and publisher). The basin walls are at ~100 km and ~1,700 km

northerly rise having actually been lifted to an elevation 0.5–1 km higher than the basin's rim. Excavation of the Caloris interior plains by impacts shows that they are underlain at a depth of several kilometers by bluer, low-reflectance material (the LRM of Sect. 9.1.3.3). The plains are not deeper in the low points, as might have been expected if the rises already existed when the flood volcanism took place, and flooded crater floors have slopes consistent with the undulation (Oberst et al. 2010; Zuber et al. 2012). The undulation therefore appears to have formed after the plains were emplaced, and also (from crater counts) after the end of the heavy bombardment.

The more northerly rise appears to be part of a broad rise that extends approximately half way around the planet at this latitude (Zuber et al. 2012), and the more southerly rise may be part of another broad rise to the south of this one. A possible origin of these rises, discussed in Sect. 9.1.3.4.3, below, is long-wavelength folding due to horizontal compressive forces in the crust.

The Northern Lowlands

The northern lowlands are an irregularly-shaped region roughly 1,800 × 3,200 km in size and ~2 km below the surrounding terrain. They include, but are not centred on, the North Pole. Although in some places their boundary follows segments of the rims of degraded impact basins, there is no evidence that the northern lowlands as a whole are impact in origin. The lowlands are mostly filled by a contiguous area of smooth plains (the northern plains) of area ~5 × 10^6 km^2, somewhat more than 6 % of the surface area of Mercury. The age of these plains, from crater counts, is comparable to that of the smooth plains in and around Caloris Basin, being emplaced near the end of late heavy bombardment, ~3.8–3.7 Gy b.p.

With no evidence of an impact origin for the lowlands and no impact basin of the right age nearby, the plains are not thought to have formed from either impact melt sheets or molten basin ejecta; rather, the most likely origin is flood volcanism involving high temperature, low-viscosity lavas, which tend to flow more quickly and over longer distances than high-viscosity flows. These plains therefore provide evidence that large-volume eruptions of lavas on Mercury did not require large cratering events (Head et al. 2011).

A broad rise ~950 km in diameter and ~1.5 km higher than the plains surrounding it is located within the plains, (Zuber et al. 2012). The surface of the rise is the same age as the surrounding plains, based on crater counts, and has the same reflectance and color. Also, the flooded floors of volcanically-buried craters on the rise all slope away from the highest point, as do the floors and rim crests of some superposed craters, indicating that the lava solidified before the terrain was uplifted. (Lava is fluid, so the surface of a lava lake within a crater will be level at the time it solidifies.) Thus, the region appears to have been uplifted after the plains formed.

If a region is simply pushed upward without lateral motion, as, for example, by the intrusion of magma under the surface or by dynamically-rising mantle plumes (the rising part of mantle convection), then the surface becomes stretched, creating extensional features such as *normal faults* and *graben* (Sect. 9.3.3.2) oriented tangentially around the rise. Essentially, the surface is pulled apart, creating a fault in which one side sinks relative to the other (a normal fault), or two parallel faults in which the surface between them sinks (graben). The observed lack of such features around the rise (Dickson et al. 2012) suggests that it is not extensional in origin. Of the various possible other origins put forward, the most likely appears to be similar to that of the undulations in the floor of Caloris Basin: long-wavelength folding due to lateral compression of the crust as the planet cooled (Dickson et al. 2012), described in Sect. 9.1.3.4.3, below.

The size-frequency distribution of craters across the northern plains is essentially uniform. Because the size-frequency distribution varied over time as the population of impactors changed, this homogeneity and that of the reflectance and color suggest that the northern plains formed over a geologically-short time interval from a single source region in the mantle. The plains therefore differ in character and composition from most mare regions on the Moon, where a series of mineralogically distinctive basalts (requiring different source regions) were emplaced over an extended period of time (Head et al. 2011).

The above observations, combined with the fact that smooth plains occupy about 40 % of Mercury's surface, indicate that there was extensive partial melting of the mantle near the end of late heavy bombardment, accompanied by widespread eruption of flood lavas with high effusion rates (i.e., gushy!).

An interesting aspect of the northern lowlands is that, unless there is a similar area near the south pole, the existence of a large lowland area close to the north pole suggests that the region may have migrated to the pole to maximize the moment of inertia about the rotation axis (Zuber et al. 2012). If this reorientation occurred, it must have happened after Mercury's outer layers had cooled enough that they had the mechanical strength to preserve the shape of the planet as the re-alignment took place.

Tectonic Features

The most common tectonic features on Mercury were not caused by dynamic mantle processes (e.g., convection), as on Venus and the Earth, but by global contraction as the planet cooled. Rock has a smaller coefficient of thermal

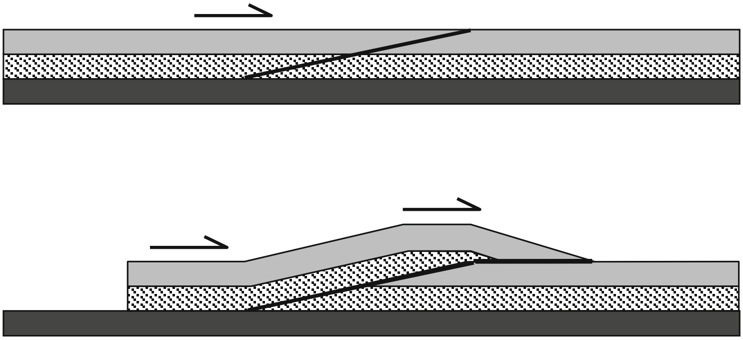


Fig. 9.7 Elevation profile of a thrust fault. Here, horizontal compressive stresses in the crust have pushed the upper two layers on the left upward along the fault and toward the right over the upper two layers to the right of the fault

expansion than does iron, so, as Mercury cooled, the rocky crust underwent a smaller amount of thermal contraction than did the iron core. However, all layers had to contract together, so the extra contraction required for the crust was made up for, at least in part, by slippage along *thrust faults*: the surface area of the planet decreased when segments of crust were thrust up and slid over the crust below (Fig. 9.7).

Three types of thrust-fault-related features have been found in images from *Mariner 10* and *MESSENGER*:

- *Lobate scarps*, or *rupes*, (Figs. 9.8 and 9.9) are the most common tectonic feature on Mercury (Watters et al. 2009). Most are asymmetric in cross-section, with a long, gentle slope on one side leading to a steep scarp (cliff) hundreds of metres to >1 km high on the other. They extend in a linear or arcuate fashion for distances ranging from tens to hundreds of kilometers; Beagle Rupes, discovered during the first *MESSENGER* flyby, is ~600 km in length and up to ~1.5 km high. The scarps often cut across pre-existing craters, producing offsets of the walls and floors in a fashion typical of thrust faults. Measured offsets (i.e., horizontal displacements of the upper material over the lower) are typically ~1–3 km; consequently, the Mercurian crust has been shortened by this amount in the direction perpendicular to the fault line. For a 1.5-km-high scarp that has been thrust 3 km horizontally, the angle of dip of the fault is $\tan^{-1}(1.5/3) = 27°$.

  Lobate scarps are found in terrain of all ages, from the old, intercrater plains (>4 Gy) to the youngest-known smooth plains (~3.5 Gy), indicating that the scarps have formed continuously over geological times rather than in a single, short-lived event (after Watters et al. 2009).
- *High-relief ridges*, or *dorsae*, the least common of the three tectonic features discussed here, are long, relatively broad landforms up to ~1 km high, and can occasionally transition into lobate scarps (Watters et al. 2009). One example

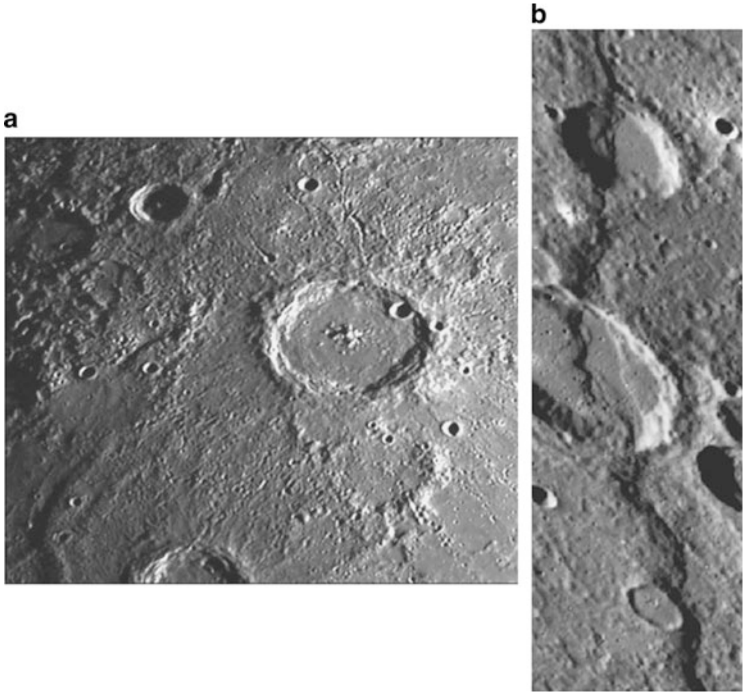


Fig. 9.8 (**a**) Large craters with central peaks and lava in-fill are not rare on Mercury. (**b**) The *Discovery Rupes,* a scarp seen at low Sun angle. JPL/NASA/Northwestern University images PIA02424 and PIA 02417, respectively

discovered in *MESSENGER* images is >600 km long and up to ~60 km wide, and transitions at its northern end into a lobate scarp that continues for almost another 400 km. The total length of the feature is thus >1,000 km. High-relief ridges can also cut across pre-existing craters, deforming the walls and floors, and are believed to be caused by reverse faults (high-angle thrust faults).

- *Wrinkle ridges* are found on smooth plains. They are lower and more complex than high-relief ridges, often having the form of a broad, low-relief arch with a narrow superimposed ridge, and are believed to form by a combination of folding and thrust faulting (Watters et al. 2009).

From measurements of the lengths of the lobate scarps, high-relief ridges and wrinkle ridges visible in areas near the terminator in *Mariner 10* and *MESSENGER* images (where they are most easily seen), occupying a total area of 16 % of Mercury's surface, Di Achille et al. (2012) found a total areal strain (decrease in surface area) of 0.24 %, assuming a thrust-fault dip angle of 30°. With the assumption that this value is uniform over Mercury's total surface area, the corresponding decrease in radius is 2.9 km.

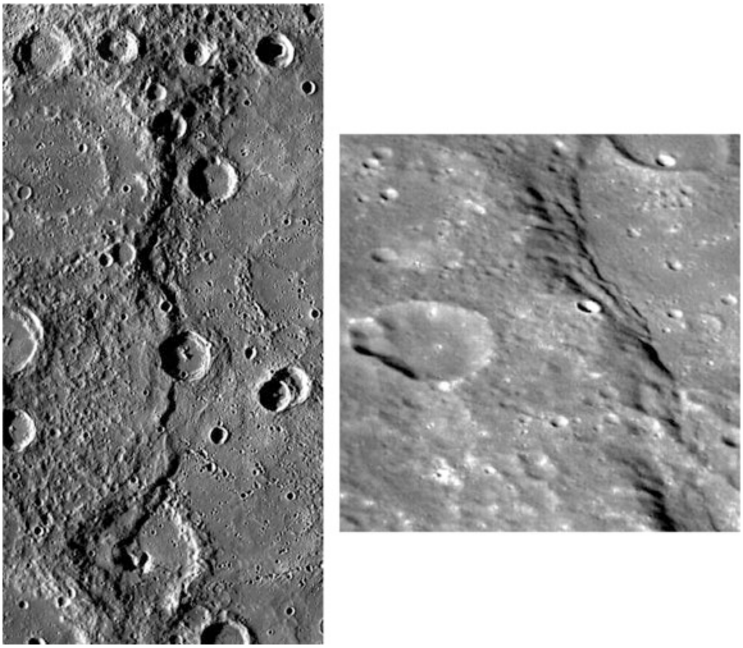


Fig. 9.9 Scarps photographed by the *MESSENGER* spacecraft. *Left*: Victoria Rupes, ~500 km in length. A large, peak-ring crater can be seen at the *top left*, with ancient lava flows having flooded its floor and almost covering the peak ring. North is up. Image PIA15614. *Right*: A detail of an unidentified scarp in the northern hemisphere, showing the structure and slope of the scarp. North is toward the *bottom right*. Image PIA15483. Both images courtesy of NASA/Johns Hopkins University Applied Physics Laboratory/Carnegie Institution of Washington

There is a problem, however, in that thermal evolution models of Mercury predict 5–6 km contraction, which is approximately twice the value found above. A possible solution (Dombard et al. 2001) is that, under compressive stress, the crust may be unstable to long-wavelength, low-amplitude sinusoidal folding. Using a planar model, Dombard et al. (2001) found wavelengths of 100–1,000 km and wavelength/amplitude ratios of ~100, reminiscent of the undulation in Caloris Basin in Fig. 9.6, the extended rises of which they appear to be a part, and the rise in the northern plains. The observed amplitudes of ~1–2 km would be sufficient to account for a significant part of the required contraction. In this model, therefore, long-wavelength rises and undulations are contractional features, rather than being caused by magma intrusion or mantle processes such as convection.

9.1.3.5 Surface Composition

Refractory Elements: Mg, Fe, Ti, Al and Ca

Although Mercury looks superficially similar to the Moon, its surface composition is different in significant ways (Nittler et al. 2011). The Mg abundance relative to Si is higher and the Al and Ca abundances are lower than either the Moon or the bulk silicate Earth (the mean composition of the Earth's crust and mantle together), giving a composition somewhat more mafic than typical basalts. The low Al and Ca abundances correspond to a low abundance of plagioclase feldspar. In fact, no examples of a plagioclase-rich crust similar to the lunar highlands have been found on Mercury.

The surface is also low in Fe and Ti, ranging from 0.5 to 3.7 wt% for Fe, and an upper limit of ~0.8 wt% for Ti (Nittler et al. 2011; Charlier et al. 2012). In the highly reducing conditions found on Mercury, Fe occurs mostly as a metal phase. However, for the purpose of comparison with other planets where oxidizing conditions prevail, these values correspond to ~1 wt% TiO_2 and 0.6 to 4.9 wt% FeO. The low FeO content may be the reason for the lack of a lunar-like plagioclase-rich crust: below a few wt% FeO in a magma ocean, plagioclase does not float to the top.

In the absence of a plagioclase-rich flotation crust, the magma ocean (if it existed) would be expected to solidify and then later be resurfaced by lava flows after local re-melting by radioactive heating in the mantle. In this case, the surface abundances would reflect the composition of the mantle source regions; i.e., the bulk silicate composition (crust and mantle) should also be < 4.9 wt% FeO (Robinson and Taylor 2001; Nittler et al. 2011).

Volatile Elements: S, Cl and K

If Mercury had formed exclusively from high-temperature material in the solar nebula at ~0.4 au from the Sun, we might expect it to have high abundances of refractory elements and low abundances of volatiles. However, as discussed above, the refractory elements Ca and Al are in low abundance; and, as discussed next, there is increasing evidence that Mercury is not depleted at least in the volatiles S, Cl and K.

The sulfur abundance measured in Mercury's surface materials by the *MESSENGER* X-ray spectrometer is much higher than that of the other terrestrial planets, ranging from 1.4 to 3.9 wt% compared to $\lesssim 0.2$ wt% for the bulk silicate Earth, lunar silicates, and stony meteorites from Mars and differentiated asteroids (Nittler et al. 2011; Charlier et al. 2012). The low S abundance on these objects is believed to result from loss of volatiles during planet formation and/or sequestration into planetary cores. Mercury's value is more in line with that of enstatite chondrite (EC) meteorites (Sect. 15.2.4), as are other aspects of its composition, suggesting that Mercury has retained its initial abundance of S. Likewise, the measured Cl abundance in Mercury's crust is $\lesssim$0.2 wt%, compared to ~0.02 wt% Cl for Earth's upper continental crust (Nittler et al. 2011).

The K/Th ratio measured by the *MESSENGER* gamma-ray spectrometer is similar to that of the other three terrestrial planets (Peplowski et al. 2011). Since K is moderately volatile and Th is not, volatile depletion would have preferentially reduced the abundance of K relative to Th and resulted in a K/Th ratio much lower than observed.

Further evidence of volatiles on Mercury is provided by the discovery in *MESSENGER* images of rimless, irregularly-shaped hollows tens of metres to several kilometers across (Blewett et al. 2011). The hollows are shallow compared to their horizontal dimensions, and generally have flat, smooth floors. Most, although not all, have high-reflectance interiors and halos; in fact, they are among the brightest features on the planet: hollows on the floor of the 97-km-diameter crater Tyagaraja have a reflectance of 0.140 at 559 nm, 2.5 times the global average of 0.057. All known hollows are associated with impact features, being found on crater and basin floors, rims, and central or other interior peaks. These features are all composed of material that has been excavated from depths of $\lesssim$2 km to ~20 km.

Several characteristics argue against a volcanic origin for the hollows, as calderas, vents or collapse pits; e.g., their shallow spectral slope ("blue" compared to the much redder color of pyroclastic deposits), distribution (impact-related terrain) and specific locations (e.g., crater central peaks, which are unlikely locations for volcanism) (Blewett et al. 2011). Rather, it is believed that the impacts have exposed volatile-rich material to the low-pressure, high temperature environment of Mercury's surface and shallow subsurface, and sublimation of the volatiles has resulted in mass wastage and subsequent collapse of the remaining material. The collapse in turn exposes fresh volatiles, maintaining the bright appearance. Dark hollows are those for which volatiles are no longer exposed and the process has ceased. Space weathering (micrometeoroid impacts and/or solar wind sputtering) can also erode the volatiles.

Identifiable spectral features for the hollows have not yet been found in the *MESSENGER* multispectral imagery, so the volatiles involved have not been identified. However, the floor of one particular 33-km-diameter crater is filled with impact melt, and the 30-m-deep upper layer of this melt is densely covered with hollows (Vaughan et al. 2012). This layer evidently formed by differentiation of the impact melt as it solidified, and may be composed of either a sulfide (MgS or CaS) or a chloride (MgCl or CaCl).

Water at the Poles

Bright radar reflections were first detected in Mercury's polar regions in 1991 using the 300-m-diameter Arecibo radio telescope as both transmitter and receiver, and also the Goldstone 10-m dish as transmitter with the Very Large Array as receiver. The region near the south pole is shown in Fig. 9.10.

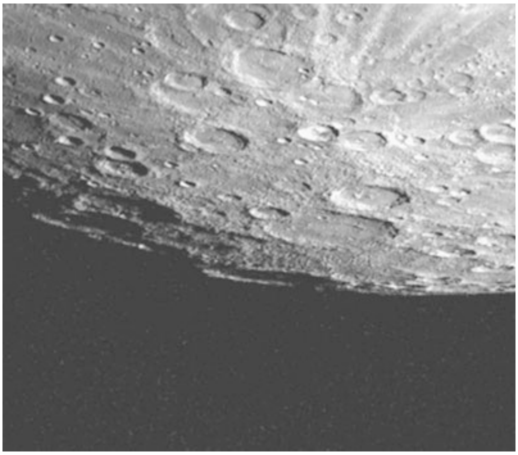


Fig. 9.10 The South Pole is in the crater with the illuminated far rim at the *extreme bottom center* of the image. Radar echoes suggest ice in the perpetually cold interiors of Mercury's polar craters. Bright crater ejecta rays are seen at the *top* of the image. *Mariner 10* image (NASA/JPL/Northwestern University, PIA02941/02415)

Subsequent observations with the newly-refurbished Arecibo telescope in 1998–2005 allowed the distribution of these radar-bright regions to be mapped with 1.5-km resolution (Harmon et al. 2011). Comparison with *Mariner 10* images suggested that these are all located in *permanently-shaded regions* (*PSR*s) within craters, and this has been confirmed by *MESSENGER* (Chabot et al. 2012, Neumann et al. 2012; Paige et al. 2012). They can also lie in the permanent shadows of poleward-facing scarps. The reflections are believed to be from water ice on the basis of polarization characteristics, radar albedo, and their apparent location in PSRs.

Water ice can remain stable for billions of years if its temperature remains <110 K (Neumann et al. 2012). However, while PSRs do not receive direct sunlight, they are subject to heating from nearby terrain (e.g., from the sun-facing wall of the crater) that absorbs sunlight and re-radiates it into the PSR at infra-red wavelengths. As a result, although the annual average temperature at the surface is ~100 K in PSRs, annual maximum temperatures at the same surface can be as high as 170 K (Paige et al. 2012); thus, exposed ice will sublime and disappear in geologically short timescales. If the ice is overlain by a layer of insulating material, however, the temperature variation is reduced. The depth of penetration of Mercury's annual temperature wave is ~0.4 m, and even ~10 cm below the surface the temperature remains low enough for the ice to be stable (Neumann et al. 2012).

A clue to the nature of the insulating material is provided by the Mercury Laser Altimeter (MLA) used by *MESSENGER* for mapping terrain. At the 1,064-nm wavelength of the laser, the surfaces of all radar-bright areas are significantly darker than the terrain around them (Neumann et al. 2012; Paige et al. 2012). This is an unexpected result if the ice is exposed at the surface, so the ice must be overlain by dark material. Impact gardening mixes material both horizontally and vertically, so the process creating these dark deposits must be on-going, and must take place at a rate fast enough to prevent impact gardening from homogenizing the dark areas with their surroundings.

Mercury is already a very dark object; in fact, few solar system materials are darker. Among those that are, the most common are the complex hydrocarbon compounds found in comets and volatile-rich asteroids. These compounds are less volatile than water, and are stable at temperatures of 150–170 K. A possible origin of the dark areas, therefore, is that they are cometary debris that has been cold-trapped in PSRs. Sublimation of the more volatile components leaves a dark, organic-rich regolith ~10 cm thick overlying and insulating the water ice below it (Neumann et al. 2012; Paige et al. 2012).

An interesting question is raised by Paige et al. (2012): “Why do all the thermal niches for water ice on Mercury appear to be filled, whereas most of the Moon’s niches appear not to be filled to nearly the same degree?” After all, subsurface temperatures in lunar PSRs are ~35 K, compared to ~100 K on Mercury. Why would a colder PSR not cold-trap a larger, rather than a smaller, reservoir of ice? This initially counter-intuitive result appears to follow from the fact that the much lower temperature of the lunar PSRs strongly inhibits diffusive migration of water into the regolith; the process is much more efficient at the warmer temperature on Mercury.

9.1.4 Interior Structure

From observations of gravity and topography, and constraints on the depth extent of thrust faults from tectonic models, the mean crustal thickness is believed to be ~50 km (Smith et al. 2012). It is thicker (50–80 km) near the equator and thinner (20–40 km) under the northern lowlands. Given the observed surface composition, which is somewhat more mafic than basalt, the mean crustal and mantle densities are ~3,100 and ~3,300 kg/m^3, respectively (Smith et al. 2012, supplemental materials). As expected from the more mafic composition of Mercury’s crust, the ~200 kg/m^3 density difference at the crust-mantle boundary is smaller than at Earth’s Moho; the density difference at the latter is ~500 kg/m^3 under continental crust and ~350 kg/m^3 under oceanic crust.

Mercury’s core, or perhaps only its outer core, is molten. Because of Mercury’s elliptical orbit, its orbital speed varies while, in principle, its spin rate remains constant. However, Mercury’s equator is slightly elliptical (Sect. 9.1.3.2), and because its rotation is locked to its mean orbital period in a 3:2 spin-orbit couple, the gravitational torque of the Sun on Mercury’s asymmetric figure varies in a repeating pattern. This periodically-reversing torque causes small, periodically-reversing deviations of the rotational period from its resonant value of 2/3 of the mean orbital period. The resulting small-amplitude oscillations in longitude relative to the longitude for a constant spin rate are referred to as *forced librations*.

Because the forced librations arise from the orbital motion, the forcing term and the rotational response both vary with an 88-day period. From radar observations, the amplitude of these forced librations is $\phi_{\max} = 38.5 \pm 1.6$ arc secs (Margot et al. 2012; see also Margot et al. 2007).

If Mercury's core is solid, then the core and the rigid outer shell (mantle + crust) are coupled together and the entire planet rotates as a solid body. The amplitude of the forced librations is then related to the moments of inertia (MoI) of the planet by

$$\phi_{\max} = \frac{3}{2}\frac{(A-B)}{C}f(e) \tag{9.4}$$

(Margot et al. 2007), where A and B are MoI along orthogonal axes in the equatorial plane, C is the polar MoI of the entire planet (core, mantle and crust), and $f(e)$ is a power series in the orbital eccentricity, e. If the core, or at least the outer core, is molten, then to an excellent approximation the rigid outer shell is decoupled from the core, and C in (9.4) is replaced by the polar moment of inertia of just the rigid outer shell, C_m.

C and C_m can be calculated from measured values of (A − B), the obliquity (Sect. 9.1.3.2), and the second-degree zeroth-order and second-degree second-order harmonics of Mercury's gravity field, C_{20} and C_{22} (Margot et al. 2012, 2007; Smith et al. 2012). Using observations of Mercury's spin state from 2002 to 2012 and values of C_{20} and C_{22} obtained from MESSENGER observations, $K = C/MR^2 = 0.346 \pm 0.014$, where K is the coefficient of the moment of inertia (Sect. 5.4.3) and M and R are the planet's mass and radius, respectively, and the ratio $C_m/C = 0.431 \pm 0.025$ (Margot et al. 2012). Comparison of the resulting values of $\phi_{\max}$ from (9.4) to the 38.5 arc-sec measured value indicates that the core, or at least the outer core, is fluid. Because of the decoupling, however, this method does not constrain the existence, let alone the size, of any solid, inner core. This point is discussed below.

Given the values discussed above for the crustal thickness, crustal and mantle densities, planetary mass and radius, and C and C_m, one can find the outer radius of the fluid core, r_C, and the mass and therefore mean density of the solid, outer shell (*c*rust + *m*antle), ρ_{CM}: $r_C = 2{,}030 \pm 37$ km and $\rho_{CM} = 3{,}650 \pm 225$ kg/m^3 (Smith et al. 2012).

With these values of r_C and ρ_{CM} and the mass and radius of Mercury, the mean density of the core is 6,740 kg/m^3. The uncompressed density of iron is 7,874 kg/m^3, so one or more lighter elements must be mixed with the iron in Mercury's core. The most likely elements appear to be S and Si: the low Fe abundance and high S abundance in Mercury's surface materials suggest that Mercury formed from highly-reduced precursor materials (e.g., enstatite chondrites), and this highly-reduced environment (low oxygen abundance) favors the partitioning of Si and S into the iron core (Smith et al. 2012).

The mantle density down to the base of the source regions for the flood basalts is ~3,300 kg/m^3, so a mean crust/mantle density of 3,650 kg/m^3 requires that the layer below these source regions be very dense. The material in this region could arise from the molten core as follows (Smith et al. 2012). At pressures below 15 GPa, the Fe-S-Si alloy would have two immiscible liquids, with the more buoyant, S-rich liquids collecting near the top of the core. Any solid FeS that forms would be more buoyant than the residual liquid, and would rise to the base of the silicate mantle, forming the dense layer required above. (The uncompressed density of FeS is ~4,600 kg/m^3.) The thickness of this solid FeS layer is not well constrained by the available data, and could be anywhere from 10 to 200 km thick.

The elliptical equatorial mass distribution of Mercury's mantle causes constant-density surfaces in the core to take on a similar shape. A solid, inner core should therefore also undergo forced librations. However, because its moments of inertia would be different from those of the mantle, its libration period should also be different, and the resulting gravitational interaction with the mantle should modify the mantle libration in observable ways (Veasey and Dumberry 2011). Although there are insufficient data to see this at the present time, another decade (from 2011) of accurate libration measurements should allow both the existence and the size of the solid, inner core to be determined.

In the meantime, the existence of a solid, inner core is strongly suggested by the following line of reasoning. First, Mercury has a magnetosphere generated by dynamo action in its core (see Sect. 11.7.1). Second, dynamo action requires that at least part of the core be both molten and convecting (Stevenson et al. 1983). (This requirement also supports the conclusion above that a light impurity is mixed with the iron in the core, because, over the age of the solar system, a planet the size of Mercury would be expected to have cooled enough that a core of pure iron would be completely solid by now. The presence of the light element lowers the temperature at which the core becomes fully solid: as pure iron solidifies from the molten core, the remaining melt becomes more enriched in the light element until the eutectic composition and temperature are reached; see Sect. 5.4.4.5) Third, convection is very efficient at transporting heat; so if the core had remained completely molten then it is likely to have cooled sufficiently over the age of the solar system that the temperature gradient would no longer be adiabatic, and convection would have ceased (refer back to Sect. 4.5.3).

If iron is freezing to form a solid, inner core, then one of at least two modes of convection can occur (Stevenson et al. 1983):

1. Thermal convection, arising from a temperature gradient: gravitational potential energy released by solid iron grains sinking through the outer core onto the inner core heats the core to an adiabatic temperature gradient.
2. Chemical convection, arising from a composition gradient: as Fe freezes onto the inner core, the residual melt is less dense than the melt above it, and convects upward while the denser melt sinks to replace it.

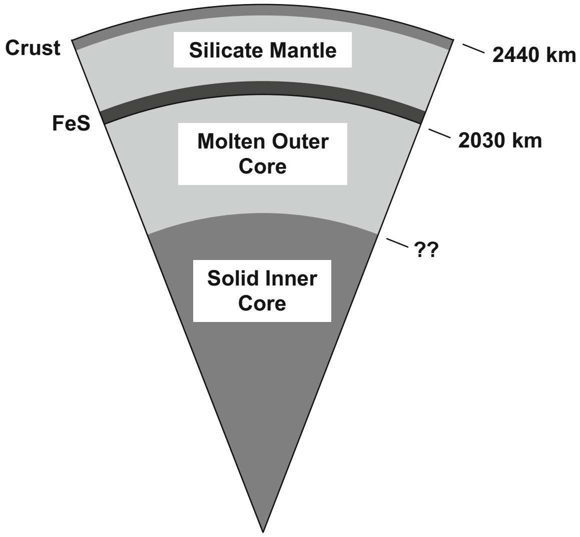


Fig. 9.11 Internal structure of Mercury. The crust is ~50 km thick, and the possible FeS layer is between ~10 and ~200 km thick. The silicate mantle is therefore between 160 and 350 km thick. The molten outer core is iron with one or more lighter elements, possibly S and Si. The solid inner core, believed to exist on several grounds, would be pure iron. The size of the inner core is unknown

The first process appears to be dominant in the Earth's core, whereas the second is more likely dominant in Mercury's core (Stevenson et al. 1983). In either case, the existence of a magnetosphere around Mercury appears to require a solid, inner core in addition to the fluid outer core, but the current observations do not constrain its size.

Figure 9.11 illustrates the internal structure of Mercury as described above.

9.1.5 Mercury's Origin

A model for Mercury's origin must include an explanation of the following properties:

1. Very high total abundance of iron. Mercury's iron-rich core extends 83 % of the way to the surface and takes up 71 % of its mass. (For the Earth, the values are 45 % and 33 %, respectively.) As a consequence, Mercury's mantle is much thinner than that of any other terrestrial planet.
2. Very low iron abundance in crust and mantle silicates (Fe is generally listed as FeO when discussing elemental abundances in silicate minerals; see Sect. 8.3). During planetary differentiation and core formation in terrestrial planets, metallic iron sinks to the centre, but iron bound in silicates remains in the crust and mantle. The high total Fe abundance and low FeO abundance therefore say something either about the primordial materials from which Mercury formed or about the processes that occurred between Mercury's formation and the present day.

Table 9.2 FeO in planetary basalts and interiors

Planet	FeO (wt%) in basalts	Bulk silicate planet[a]
Mercury	<4.9	<4.9
Venus	8.6	7–8
Earth	10.5 (MORB[b])	8
Mars	20	18
Vesta	18.5	20

[a] Mean composition of the planet's crust and mantle together
[b] Mid-ocean ridge basalts

3. High abundance of Mg and low abundances of Al and Ca. Mercury is therefore depleted in plagioclase relative to the Earth.
4. A higher abundance of the volatile elements S and Cl than for the Earth, and a K abundance similar to that of Earth.

Prior to *MESSENGER*, it was often assumed that Mercury would be strongly depleted in volatiles relative to the Earth, as a consequence of having formed in a high-temperature region of the solar nebula where volatiles would not condense. Mercury's thin mantle needed explaining, and two popular models were either a giant impact that ejected a large fraction of the mantle, or large-scale evaporation of Mercury's surface by the early Sun. (However, see Taylor and Scott (2001) for arguments in the pre-*MESSENGER* context against these two processes, and in favor of formation from locally available materials.)

Both a giant impact and large-scale evaporation would have eliminated a plagioclase crust and left mantle material exposed, accounting for points 1 and 3, above. They also would have evaporated a large proportion of any volatiles present. With *MESSENGER*'s discovery of abundant volatiles at Mercury's surface and down at least as far as the mantle source regions, both models are now seen as unlikely (Nittler et al. 2011). The formation of Mercury primarily from high-temperature, refractory-rich and volatile-poor material is also ruled out.

The current thought is that Mercury's present composition reflects the precursor materials from which it formed. A clue may be provided in Table 9.2, adapted from Table 2 of Robinson and Taylor (2001) and updated with the *MESSENGER* results for Mercury from Nittler et al. (2011). The FeO abundances in this table show a gradient from ~20 wt% for Mars and Vesta to 7–10 wt% for the Earth and Venus to <5 wt% for Mercury. Thus, Mercury's FeO abundance can be seen, not as abnormal, but as arising from a composition gradient that was present in the solar nebula.

Possible formation scenarios include (1) a mixture of refractory-enriched material and material with a composition similar to the Earth, or (2) highly reduced and/or metal-rich chondritic meteorite compositions (enstatite and/or CB chondrites, respectively); see Nittler et al. (2011) for references. Enstatite chondrites, which formed under highly reduced conditions (low oxygen abundance), provide a reasonably close if imperfect match, including a sulfur enrichment at about the same level as observed on Mercury.

Thus, evidence indicates that Mercury formed preferentially from highly reduced, but not strongly volatile-depleted, precursors with compositions similar to enstatite chondrites. A difficulty is that enstatite chondrites cannot provide the very high total Fe abundance relative to Si observed in Mercury, but it is also probable that not all of the primitive building blocks from the solar nebula survived to the present day, to be sampled by astronomers. Highly-reduced objects with higher metal abundances than enstatite chondrites may well have existed in the early inner solar system (Nittler et al. 2011).

9.2 Venus

9.2.1 Visibility and General Properties

Venus sometimes appears as the brightest object in the sky after the Sun and Moon, achieving −3.3 to −4.4 V magnitude at ~39° elongation (not quite at maximum elongation: 48°). As either a morning or an evening star, it commands attention.

The planet was considered a manifestation of a god in many past cultures (e.g., as Ishtar/Inanna in the Middle East, Hesperus in its evening star manifestation and Phosphorus [Lucifer, bearer of light] in the morning, further West). In Mesoamerica, it was considered a fearsome god, and the spilling of blood was required to placate it. For extensive discussion of the perception of Venus from culture to culture, and its configurations in the sky, see Kelley and Milone (2011, Sections 2.4, 3.15, 5.4, 7.1.4, . . .).

The bulk properties of the planet are summarized in Table 9.1. The elements for an osculating orbit at the epoch JDN 2456310.5 = January 19, 2013 at 0^{h} U.T. are as follows:

$$\begin{aligned}
&a = 0.72333\ \text{au};\\
&e = 0.00678;\\
&i = 3\overset{\circ}{.}39;\\
&\Omega = 76\overset{\circ}{.}6439;\ \text{and}\\
&\varpi = 131\overset{\circ}{.}864.
\end{aligned}$$

Its orbital eccentricity is smaller than that of the Earth, so its orbit is more circular, and, with the small inclination, it lies close to the ecliptic plane: $i = 3\overset{\circ}{.}39$.

Its period of revolution is $224\overset{d}{.}7$, so that its synodic period is $1^{y}219^{d} = 1\overset{y}{.}599$.

Transits of Venus are observable when inferior conjunctions occur near a node passage. The first observation of a Venus transit was by Jeremiah Horrocks (predicted from his amended form of Kepler's 1627 Rudolphine Tables) and William Crabtree, who observed from a separate location, on November 24, 1659. The most recent transits occurred on June 8, 2004, and June 5–6, 2012, and before that, on December 9, 1874, and December 6, 1882; the next will occur on December 11, 2117, and December 8, 2125. Note the pairing of transits in short

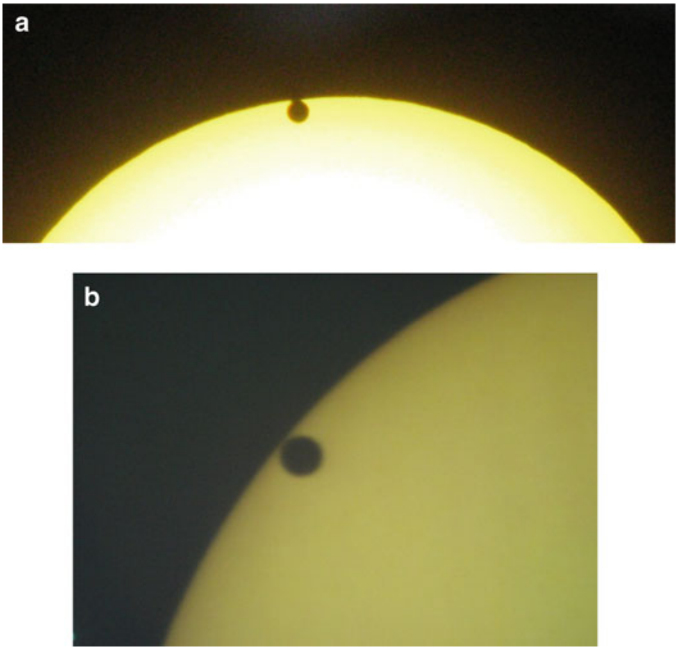


Fig. 9.12 The black drop effect. (**a**) As imaged on June 5, 2012, at 5:23 p.m. from a low-elevation site in Lake-in-the-Hills, IL, showing the black drop effect between the silhouette of Venus and the solar limb. Taken with a Canon PowerShot SD100 Camera, ISO-400 speed, 1/400 s exposure time, and f/2.8 f-stop. Courtesy B. V. L. Milone. (**b**) The transit viewed through the 10-m Keck telescope at 4,200 m elevation on Mauna Kea, not showing the black drop effect. Photograph of a viewing screen at the Rothney Astrophysical Observatory, by E. F. Milone

intervals, and the change of node across a much larger interval. Such events were sought to determine the scale of the solar system, as soon as ephemerides could be accurately computed—following Johannes Kepler. Edmund Halley writing in 1716 argued passionately for observations of the transit in order to determine the au precisely. His method involved the precise timings of the second and third contacts, when the disc of Venus is just within the disc of the Sun. Unfortunately, the "black-drop" effect (see Fig. 9.12) prevented the determination of the times of contact to the high precision that Halley had anticipated. The drop extends the time interval when the edges of the discs appear to be in contact, thus precluding the recording of a precise instant. The images of the 2012 transit from a low-elevation site (Lake in the Hills, IL, near Chicago) and from the Keck Telescope on Mauna Kea (4,200 m) on the Island of Hawaii, where the drop effect was not apparent, suggest that site elevation, with its accompanying atmospheric refraction and scintillation, is a factor in the phenomenon; but see Pasachoff et al. (2003) for a fuller discussion.

The distance rotated through by the Earth in the course of the transit can be used as a baseline for a single observer (Method 1 in Fig. 9.13). Alternatively, the

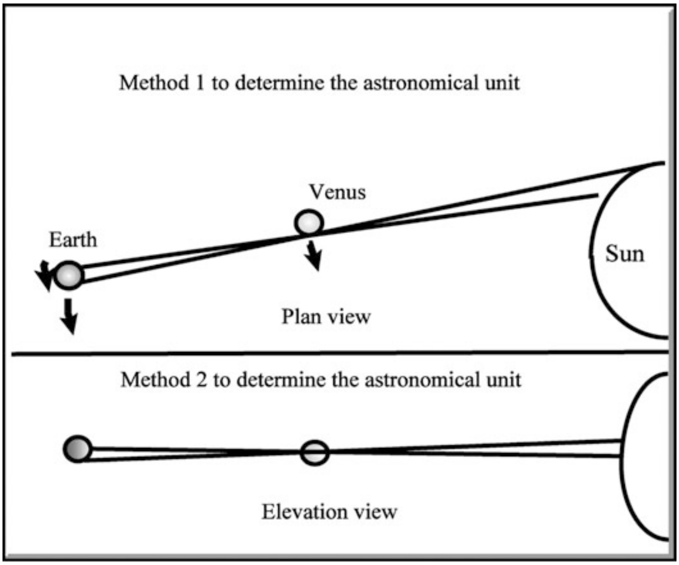

Method 1 to determine the astronomical unit
Venus
Earth
Sun
Plan view
Method 2 to determine the astronomical unit
Elevation view

Fig. 9.13 Use of the transit of Venus to find the distance of Earth to Venus in kilometers, and thus the astronomical unit

known separation of two observatories on the Earth can provide the baseline of a triangle, the apex of which is at Venus (e.g., Method 2, in Fig. 9.13). The relative timings of the immersion and emersion and/or the exact placement of chords across the Sun provide the necessary data. With either method the Earth-Venus distance can be found in kilometers. Setting this distance equal to the fraction of the astronomical unit expected from celestial mechanics, the value of the astronomical unit is found:

$$r\ (\mathrm{km}) = f\ a\ (\mathrm{km}) \tag{9.5}$$

where r is the determined distance, a, the astronomical unit and f, the fraction of the au represented by the distance.

Venus, like Mercury, has no natural satellite. From n-body integration work, however, its mass is found to be 0.815 that of Earth. By direct measurement before and after inferior conjunction, its radius is 6,052 km, so that its mean density is 5,200 kg m^{-3}. This implies a large iron core (see below). Yet, Venus has no detectable magnetic field unrelated to the solar wind. An internally-generated magnetic field requires that the core be both molten and convecting, and convection in a terrestrial-planet core requires the presence of a solid, inner core (Sect. 9.1.4). Venus, with a diameter and mass similar to the Earth, is likely to have a liquid core, as the Earth has; so the absence of a magnetic field suggests that Venus lacks a solid, inner core. This situation may arise because Venus' central pressure is ~20 % lower than that of the Earth, and its central temperature is slightly higher, in part because of the higher surface temperature (Stevenson et al. 1983).

The mean gravitational acceleration on the surface of Venus at the equator is 8.96 m/s^2, making it the most Earth-like planet in terms of weight on the surface. Aside from the physical measurements, however, the environment on the surface of Venus is unmatched on Earth, except maybe *in* active volcanoes!

As for Mercury, the rotation period has not been easy to measure, but here the dense clouds that veil the planet are the cause. In 1890, Giovanni Schiaparelli (1835–1910) suggested that $P_{\mathrm{rotation}} = P_{\mathrm{rev}}$, and Percival Lowell (1855–1916) concurred in 1911. In 1921, Edward C. Pickering (1846–1919) produced a value of 68$^{\mathrm{h}}$ about an axis in the plane of the orbit; W. H. Steavenson in 1924 confirmed this axis, but found a period of 8$^{\mathrm{d}}$. In 1927, Frank E. Ross (1928) proved from UV-emulsion photography carried out with the 60- and 100-in telescopes on Mt Wilson that so short a period as 68$^{\mathrm{h}}$ could not satisfy his data (references to earlier work are cited in his paper), but he was unable to find a period with much precision because of the elongated nature of the markings and their changing appearance from day to day, and settled on a value ~30$^{\mathrm{d}}$. In 1956, J. D. Kraus claimed that radio signatures had yielded a period of 22$^{\mathrm{h}}$ 17$^{\mathrm{m}}$ ± 10$^{\mathrm{m}}$. Radar determinations in 1962 finally produced an accurate result: $-243^{\mathrm{d}} = -0\overset{\mathrm{y}}{.}665 = -1.50P_{\mathrm{rev}}$. The negative period signifies that Venus has a retrograde rotation; unlike the Earth, it rotates CW as viewed from above the North Ecliptic Pole. Because the synodic period of Venus is $1\overset{\mathrm{y}}{.}60$, Venus is close to an orbital lock-in with Earth (Sect. 3.7.3). Venus appears to be less flattened than the Earth or Mars, with a potential coefficient $J_2 = 2.7 \times 10^{-5}$.

The conditions on the surface of Venus are beyond "harsh!" The Sun would be perceived only dimly through the super-dense cloud decks, and the greenhouse effect produces a surface temperature of ~750 K, under a pressure of ~90 bar at the surface. The atmospheres of Venus and the other terrestrial planets will be compared in Milone and Wilson (2014, Chap. 10), so we defer further discussion of Venus' atmosphere until then except to note that Venus' atmosphere was discovered and remarked on by the Russian scientist Mikhail Lomonosov (1711–1765) during the 1761 transit of Venus (Shiltsev 2012; but see also Pasachoff and Sheehan 2013).

Next we describe the features on Venus, known only through radar observations and mainly from the *Magellan* mission of the 1990s. Other data will be discussed in the sections to follow.

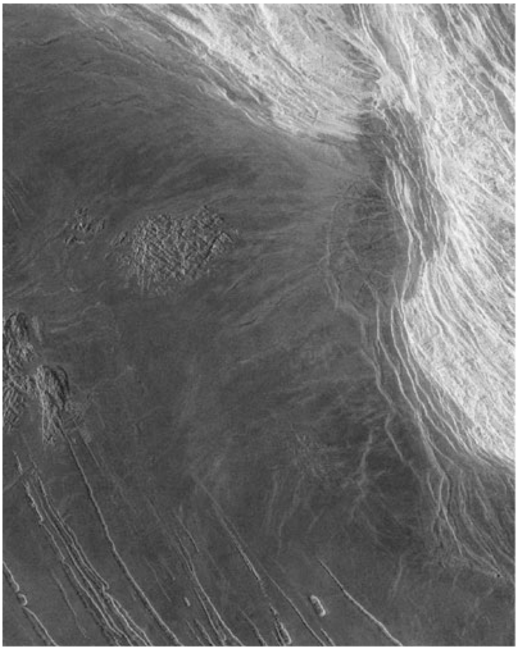
9.2.2 Types of Surface

9.2.2.1 Division in Terms of Elevation

Highlands:	8 % of Venus' surface area
Rolling plains:	65 %
Lowlands:	27 %

Venus has relatively little relief compared to the Earth or Mars (Sect. 9.1.3.3), and the frequency distribution of its features' altitudes is unimodal (i.e., one peak near

Fig. 9.14 The (*dark*) eastern edge of Lakshmi Planum and the western edge of Maxwell Montes, at 11.5 km, the highest region on Venus. At left center are tesserae, jumbled terrain due to intersecting graben. Other graben at bottom. A *Magellan* NASA/JPL image, PIA00241



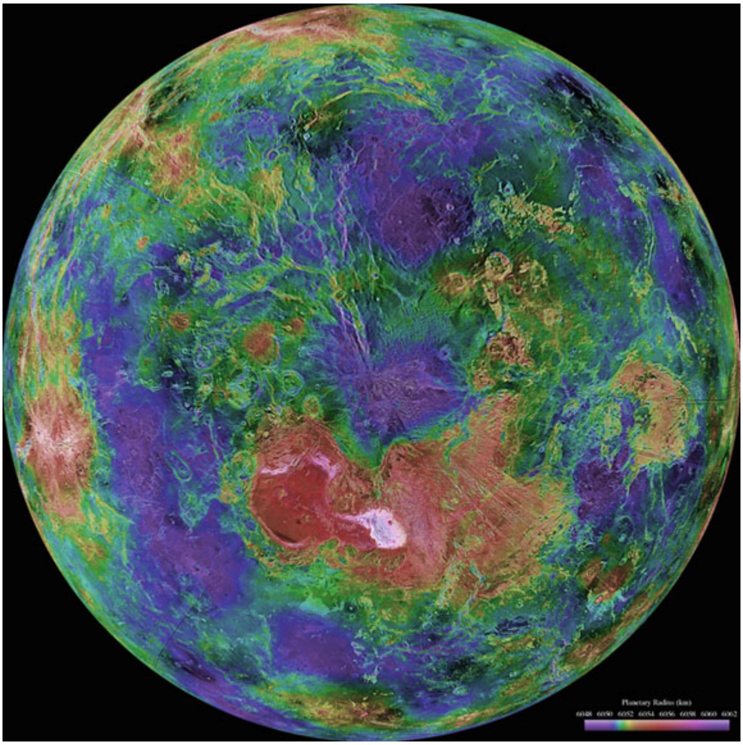


Fig. 9.15 A contour map of the northern hemisphere of Venus, based on *Magellan* radar data. Lakshmi Planum and Maxwell Montes are seen just below the center. NASA/JPL image PIA00007

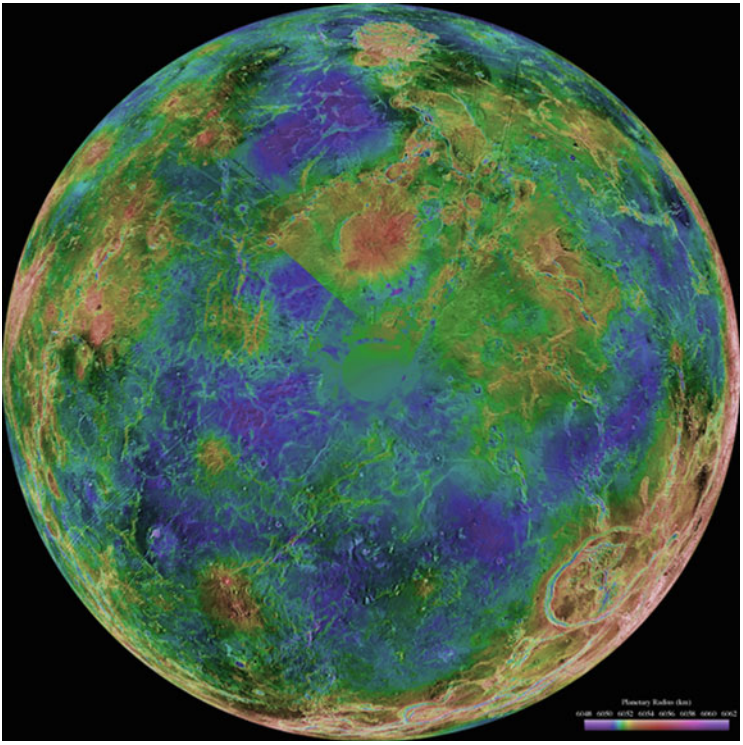


Fig. 9.16 A contour map of the southern hemisphere of Venus, constructed from *Magellan* radar data. The extensive bands on the lower right limb are features of an equatorial rift zone. NASA/JPL image PIA00008

~0 altitude, compared to Earth's distribution which is bimodal because of the distinction between continents and ocean floors, and has much greater variation); its highest mountain range, Maxwell Montes, however, is higher than the Himalayas. See Figs. 9.14, 9.15, and 9.16, and, for further discussion, Head and Basilevsky (1999).

9.2.2.2 Division in Terms of Geologic Origin

Volcanic units (lava plains):	>70 %
Highly deformed tectonic[3] units:	25 %

[3] The word "tectonic" refers to any geologic process which involves the movement of solid rock. Large-scale tectonics in the crust generally is caused by processes in the underlying mantle. In plate tectonics on the Earth, solid lithospheric plates slide around on the surface in response to convection in the mantle. Plate tectonics does not seem to occur on Venus, but other tectonic processes do.

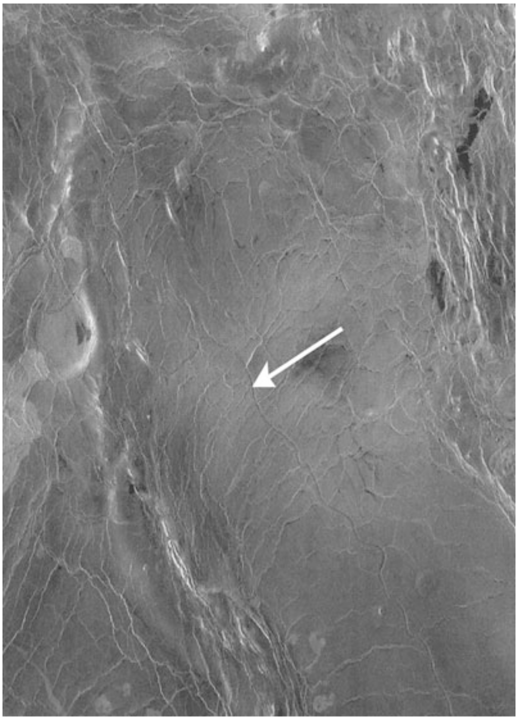


Fig. 9.17 A 600 km portion of the longest channel detected on Venus. Discovered by the orbiters during the *Venera 15–16* mission, it is more than 7,000 km long and ~1.8 km wide. A common type of feature on Venus, it is thought to be a lava channel, evidence that at least some types of lava on Venus are less viscous than Earth lavas. *Magellan*/NASA/JPL image PIA00245, with added marker

The extensive lava plains on the surface of Venus may be due partially to the high temperature environment and partially to dissolved gases within the lava, which may make it more fluid. On Earth, steam and other dissolved gases are released from the molten material at the surface, but on Venus, the 90 bars of surface pressure are sufficient to keep these gases in solution, resulting in a smoother flow than is found on Earth (Head and Basilevsky 1999). On the other hand, in another interpretation the "pancake domes" have been attributed to thick, viscous lava, which formed a cap that blocked further lava flow. There is some evidence, however, that at least some flows on Venus are less viscous than Earth lavas (note the arrow-marked, very long flow channel in Fig. 9.17).

9.2.3 Major Geologic Features of Venus

9.2.3.1 Ages of Geologic Features

Nearly a thousand (~970) impact craters have been found on Venus, ranging in diameter from ~1.5 km to 270 km. These craters provide a global average surface age of ~(750 ± 350) million years (McKinnon et al. 1997), but individual terrain features are difficult to date because of the small number of local craters. The crater

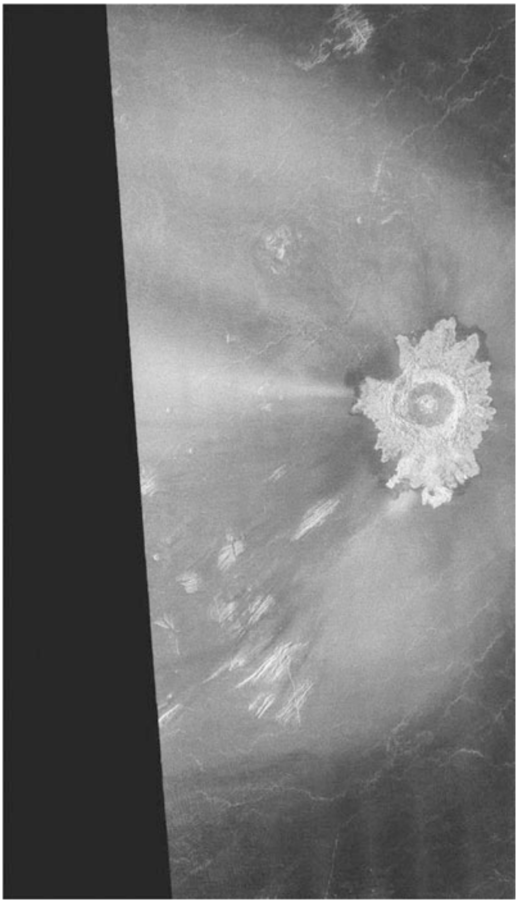


Fig. 9.18 Wind-blown deposition streaks at the Adivar impact crater on Venus as seen by *Magellan*. NASA/JPL image, PIA00083

size-frequency distribution (number of craters per unit area as a function of crater size) is an important factor in obtaining accurate age estimates, but Venus' dense atmosphere systematically biases against smaller craters: smaller potential impactors are destroyed before hitting the ground.

The crater distribution over the surface is indistinguishable from random (Turcotte et al. 1999), but the absence of markedly high- or low-crater-density regions at least suggests that the ages of individual terrains may not differ markedly from the global mean.

The lack of water on Venus means that erosion is much less important than on Earth. In fact, with only a few exceptions, and despite ages of up to ~750 million years, the impact craters on Venus appear to be in pristine condition; e.g., Adivar impact crater in Fig. 9.18 shows wind-blown deposits downwind from the crater, but the crater itself has not been visibly eroded. (Wind-blown deposits are not common in *Magellan* images, but there are some; another is seen in Fig. 9.19, downwind from a volcanic crater.) Any impact craters not in pristine condition

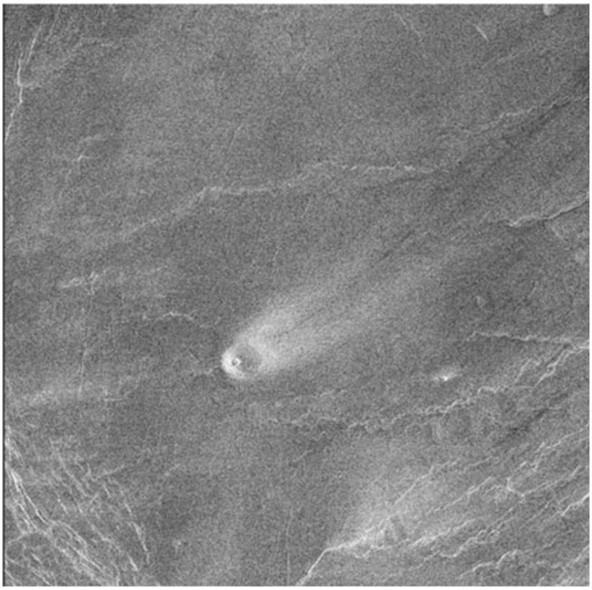


Fig. 9.19 A 35-km long deposition streak of radar-bright material on a volcano in the Parga Chasma region of Venus. A *Magellan* NASA/JPL image, PIA00243

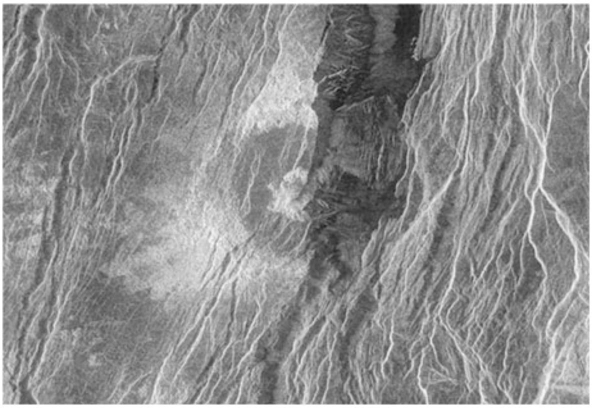


Fig. 9.20 The Summerville impact crater, 37 km diameter, half obliterated by faults in a rift area of Beta Regio. Part of the central peak region can be seen sliding into the chasm. A *Magellan* NASA/JPL image, PIA00100

have been modified by tectonic processes (Fig. 9.20) rather than by erosion. Thus, wind and chemical processes also do not cause visible erosion even over geologically-long timescales, and the extent of erosion cannot be used to estimate ages.

Although the essentially-random crater distribution provides very little information about the relative ages of different surface features, the relative ages of adjoining features can be found from stratigraphy: younger materials overlie or embay older materials. Several of these features are described below, with relative ages for the different terrains.

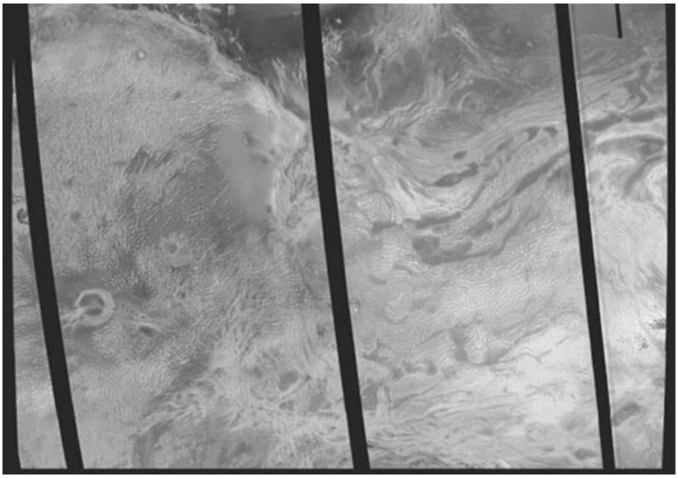


Fig. 9.21 Ovda Regio, a highland region of Venus in western Aphrodite. It rises ~4 km above the low-lying plains. It is a region of mountain belts, domes, ridges, and valleys. A *Magellan* NASA/JPL image, PIA00146

9.2.3.2 Highland Areas

Terrae: The largest highland areas on Venus. Three terrae are recognized: Ishtar in the northern hemisphere, Aphrodite near the equator, and Lada at far southern latitudes. A terra can include several of the features listed below.

Regiones: (singular: *regio*) Large highland areas that are smaller than terrae. With *Magellan* data, regiones are now known to fall into two different categories having different origins: crustal plateaus and volcanic rises. These features are described next.

Crustal plateaus: Roughly-circular, steep-sided, flat-topped upland areas ~1,500–2,500 km in diameter and ~0.5–4 km above the surrounding plains. They are believed to be supported by thickened crust or lithosphere, rather than by an on-going, dynamic process such as a mantle plume. The surface of a crustal plateau is characterized by tessera terrain, described below.

There are seven crustal plateaus on Venus: eastern Ovda Regio, western Ovda Regio, Thetis Regio (Ovda Regio is shown in Fig. 9.21; it and Thetis Regio are parts of Aphrodite Terra), Fortuna Tessera (the eastern part of Ishtar Terra), and Alpha, Tellus and Phoebe Regiones. (Phoebe Regio is unusual in that it shows characteristics of both a crustal plateau and a volcanic rise; see Hansen (2007) for more discussion.)

Volcanic rises: Roughly-circular, dome-shaped regions ~1,500–2,500 km in diameter and ~1–3 km above the surrounding plains. In contrast to crustal plateaus, volcanic rises are marked by a gentle slope outward from the centre, merging smoothly with the surrounding plains, and radial lava flows. They are believed to be supported by deep mantle plumes rising under a thick lithosphere.

The regiones associated with volcanic rises are Atla, Beta, Bell, Dione, eastern Eistla, central Eistla, western Eistla, Imdr, Laufey and Themis.

(continued)

(continued)

Volcanic rises are divided into three morphological classes (Hansen 2007):

1. Rift-dominated rises (two: Atla and Beta) are dominated by three large rift zones that form a triple junction within the rise, large shield volcanoes, but few coronae. (Rift zones and coronae are described below.) Each rift zone is 50–100 km wide, ~2 km deep and thousands of km long, and extends far beyond the boundary of the rise. [Phoebe Regio can also be viewed as a rift-dominated rise, but also has characteristics of a crustal plateau (Hansen 2007)].
2. Volcano-dominated rises (five: Bell, Dione, western Eistla, Imdr and Laufey) are dominated by one or more large volcanoes (>300 km diameter), with only minor amounts of extension; e.g., no large rift zones.
3. Corona-dominated rises (three: eastern Eistla, central Eistla and Themis) are dominated by three to eight coronae, with only minor amounts of extension. The coronae are sources of abundant volcanism.

In addition, the most prominent topographic feature of Lada Terra is Lada rise, which appears to combine the characteristics of corona-dominated and rift-dominated rises: two large coronae (850-km diameter Quetzalpetlatl and ~300-km diameter Boala) and two or perhaps three rift (graben) zones that converge on Boala Corona.

9.2.3.3 Terrain Units

Much of the information on Venus' terrains given below is from Ivanov and Head (2011), to which the reader is referred for more information. The geological map presented by Ivanov and Head covers the surface of Venus from 82.5° N to 82.5° S (89.9 % of the surface area of Venus), and the percentages listed below give the fraction of this area taken up by the terrain unit described. There is a slight uncertainty in these numbers in that 6.2 % of the mapped area consists of gaps, for which the terrain is unknown.

Tesserae: (7.3 %) Terrain with an appearance reminiscent of a parquet floor. ("Tessera" is from Greek, meaning "tile.") Tessera terrain is heavily deformed tectonically, with the basic structure consisting of extensional features (parallel graben and/or fractures) crossing compressional features (parallel ridges or folds) orthogonally or diagonally, or sometimes in a chevron or chaotic fashion. Often there are several sets of intersecting structures.

Tesserae are distributed non-uniformly over the surface of Venus, with very few occurring south of latitude 30° S. Most are roughly equidimensional or slightly elongated, with areas from ~200 km^2 to the largest, Ovda Regio, with an area of 9×10^6 km^2. Individual ridges are typically a few hundred meters to as much as several tens of kilometers in width, several hundred meters in height, and up to several hundred kilometers in length.

Tesserae are the oldest terrain in any given area of Venus, but it is not clear whether they represent a single, global terrain unit of uniform age, most of which has become covered by lava plains, or if different tesserae formed independently at different times. There is also debate on whether the compressional or extensional features occurred first, or if they formed synchronously.

(continued)

(continued)

In addition to being the terrain of crustal plateaus, tesserae also occur as inliers in volcanic plains at elevations as low as ~2 km below the mean planetary radius. These may be outcrops of old crustal plateaus that have collapsed and been partially covered by the lava of the volcanic plains.

Planitiae: Large, lowland areas. Their surface is dominated by lava plains of various types, some of which are described below, but they can also contain tectonic units such as tessera inliers, ridge belts and coronae.

Densely-lineated plains: (1.6 %) Small plains units, usually only tens of km across, that have been densely deformed tectonically by roughly-parallel lineaments or fractures. They are embayed by all other plains units that contact them, indicating that they are the oldest plains terrain unit on Venus. They are also slightly elevated, which suggests that they appear only where adjacent plains units are thin enough not to cover them. Usually, they do not contact tesserae, but in the few places that they do, they embay the tesserae. This suggests that they are younger than tesserae, but an overlap in age cannot be ruled out. In shape they can be equidimensional, elongate or arcuate.

Mountain belts: (0.3 %) Mountain ranges on Venus occur only as belts surrounding Lakshmi Planum in Ishtar Terra. Four mountain belts are recognized: Danu Montes, Akna Montes, Freyja Montes and Maxwell Montes. Each belt is ~100 to several hundred km wide and up to ~1,200 km long, and densely packed with ridges 5–15 km wide and tens to a few hundred km long. These ridges rise ~1–8 km (the highest is Maxwell Montes) above the level of Lakshmi Planum, or ~5–12 km above the mean planetary radius.

Ridged plains: (2.1 %) Volcanic plains tectonically deformed by a system of roughly-parallel compressional ridges, each typically 5–10 km wide and tens of kilometers long. In many cases the ridges occur in densely-packed belts (*ridge belts*) several tens to a few hundred km wide and hundreds of km long. Ridged plains and ridge belts appear to be similar in age to, or perhaps slightly younger than, densely-lineated plains, and older than shield plains.

Groove belts: (8.1 %) Belts a few hundred km wide and from hundreds to thousands of km long, densely packed with graben or fractures from several hundred metres up to ~2 km wide and up to several tens of km long. They tend to occur on somewhat elevated terrain, and can cut into tesserae, densely-lineated plains and ridged plains but are embayed by shield and regional plains; thus, they are younger than the former and older than the latter.

Shield plains: (17.4 %) Large plains units, thousands to millions of square kilometers in area, with abundant shield volcanoes 1–20 km diameter distributed roughly uniformly over their surface. The volcanoes are the sources for the plains, which appear to be tens of metres or less in thickness and often have a lace-like appearance. The shield volcanoes can occur in small clusters, giving the plains a hilly appearance, but this clustering is distinct from the shield clusters described below. Shield plains are younger than densely-lineated plains and older than regional plains. Tectonically, they are relatively undeformed, showing only some wrinkle ridges and a scattering of fractures or graben.

Most terrain units from shield plains onward were volcanically emplaced and show little tectonic deformation compared to earlier terrains. The stratigraphy therefore suggests a transition from a tectonically-dominated regime to a volcanically-dominated regime around the time of the emplacement of the shield plains.

(continued)

(continued)

Regional, or wrinkle-ridged, plains: (40.3 %) The most common plains unit on Venus, characterized by a smooth surface deformed by a network of sinuous wrinkle ridges formed tectonically by contraction. Individual wrinkle ridges are typically a few km wide and tens of km long. Regional plains have a basaltic composition, but no source vents have been found and their origin is unknown.

Shield cluster: (0.7 %) A tight cluster of small shield volcanoes in an area of ~1,000–20,000 km^2. This differs from shield plains, which have a lower density of shield volcanoes over an area of 1,000–10^6 km^2. The underlying surface of a shield cluster is similar to that of shield plains, but in contrast to shield plains, small lava flows from the volcanoes in a shield cluster are often superimposed on the plains surface. Also in contrast to shield plains, these flows appear stratigraphically younger than regional plains where the two are in contact.

Smooth plains: (2.3 %) Smooth, featureless plains without tectonic deformation. Most are tens of km across, but a few reach a few hundred km. They formed after the regional plains, but their statigraphic relationship to lobate plains is unclear. Some appear to be volcanic in origin, while others are associated with impact ejecta.

Lobate plains: (8.3 %) Smooth surface with bright and dark flow-like features tens of km wide and up to several hundred km long. They tend to be equidimensional, from tens of kilometers up to 1,000 km across, and often occur in the vicinity of large volcanic centers, usually in association with the large dome-shaped rises, e.g., Beta, Eistla, and Atla Regiones and Lada Terra. They appear to be outflows resulting from multiple, massive eruptions.

Rift zones: (5.0 %) (aka *chasmata*, singular *chasma*.) Similar in structure to groove belts, but on a much larger scale: rift zones can be up to a few hundred km in width and several thousand km in length. They consist of fractures and graben, each of which can be up to tens of km wide and hundreds of km long. They are found in close association with lobate plains, and therefore occur preferentially within regional highs in the equatorial zone: eastern Aphrodite and the BAT region (Beta, Atla, and Themis Regiones), with smaller numbers in Eistla Regio and Lada Terra. Rift zones and lobate plains appear to be approximately contemporaneous. See Fig. 9.20.

As an example, one of the largest rift zones, Parga Chasma, extends for 10,000 km from Atla Regio southeast to Themis Regio, and has numerous branches at angles to the main rift system. Individual troughs are discontinuous, with depths of 0.5–2 km. Including branches, the width of the system ranges from 60 to 590 km.

9.2.3.4 Other Tectonic Units

Coronae: Roughly-circular volcano-tectonic structures consisting of various combinations of a raised rim, a trough inside the rim, and an interior that can be dome-shaped, flat (forming a plateau), or a depression. They are one of the dominant tectonic features on Venus, with 513 identified (Glaze et al. 2002). Artemis Corona (Fig. 9.22), on the south side of Aphrodite Terra, is the largest, with a diameter of ~2,600 km; the remainder range in diameter from 60 km to Heng-o Corona at 1,060 km. Three occur on Ishtar Terra and five on tesserae. Most of the remaining 505 are located along chasmata, but ~25 % occur as isolated coronae in the plains and ~10 % are found on volcanic rises.

(continued)

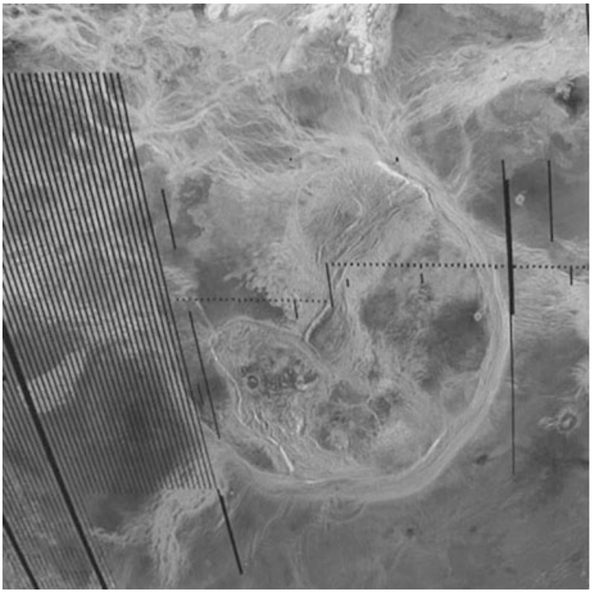


Fig. 9.22 Artemis Corona, the largest corona found on Venus. The dark linear striations are artifacts of the recording and transmission process from the *Magellan* spacecraft. NASA/JPL image, PIA00101

(continued)

Not all of the topographic features listed above (rim and trough with an interior dome, plateau or depression) necessarily occur together in each corona. In this respect, nine groups have been defined, with the third group subdivided (Stofan et al. 1997; see Table 1 of Hoogenboom and Houseman (2006) for sketch vertical cross sections):

1. Dome only.
2. Plateau only.
3. (a) Rimmed plateau. (b) Rim, inner high.
4. Rimmed depression.
5. Outer rim, trough, inner high.
6. Outer rim, trough, depression.
7. Rim only.
8. Depression only.
9. Chaotic/none.

Coronae typically contain numerous small volcanoes (<50 km diameter), with extensive lava flows both internally and externally. In fact, in the absence of plate tectonics, volcanism in coronae has been suggested as one of the main mechanisms of heat flow through the Venusian crust. There can also be a number of tectonic features: most have a partial or complete annulus of closely-spaced fractures or ridges, most commonly located on the raised rim but occasionally inside or outside the rim, and there can also be radial fractures and concentric fractures or ridges.

There is no consensus yet on how coronae form, or even if coronae at different locations form differently. Moreover, different-looking coronae may form by the same mechanism, but be at different stages of development; e.g., domes may evolve into depressions. Proposed models include:

1. Upwelling over a hotspot (Squyres et al. 1992). With some variation, the basic model begins with a rising diaper (Fig. 9.23) that initially causes up-doming of the

(continued)

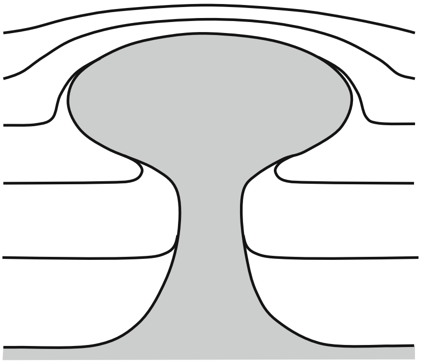


Fig. 9.23 An example of diapirism: Magma rises through cracks in the layers of surrounding rock due to buoyancy resulting from the density contrast between the magma and the rock. It forms a diapir when it reaches a level of neutral buoyancy at some depth below the surface where the downward gravitational force on the magma and the net upward force from the rock balance

(continued)

surface with accompanying radial fracturing and volcanism. The diapir impinges on the underside of the lithosphere and spreads horizontally, raising the edges of the dome and allowing the centre to relax downward to form a plateau. As the diapir cools, its density increases and its buoyancy decreases, removing the thermal support for the plateau. The interior of the plateau then relaxes to form a depression while the topography outside the edge of the plateau relaxes to form a trough with a raised outer rim.

2. Upwelling with delamination of the lower lithosphere (Smrekar and Stofan 1997). A hotspot of finite duration in the mantle creates a rising plume of hotter mantle material. The plume head encounters the lithosphere, causing up-doming of the lithosphere and surface, then spreads out under the lithosphere, cooling and increasing in density until it sinks back into the mantle in a ring around the edges of the dome. The ductile lower lithosphere becomes entrained in this outward flow, increasing the thickness of the lithosphere near the edges. When this lithosphere becomes too thick, its lower part begins to separate (delaminate) and is subducted, and an annular depression forms above it; meanwhile, the hotspot and plume die away. With the thermal support removed, the interior of the dome sinks while the ring of subducting lithosphere migrates toward the centre, along with the annular depression above it. The result is an interior depression with an elevated rim. Variations in conditions (e.g., viscosity, lithospheric thickness) appear capable of producing most of observed corona forms.

Rayleigh-Taylor instability (Hoogenboom and Houseman 2006). A Rayleigh-Taylor instability occurs when a denser fluid layer lies above a less dense fluid layer; a small perturbation can cause the two layers to overturn. In this model, a lighter crust overlies a denser lower lithosphere, which in turn overlies a lower-density layer in the upper mantle (e.g., this layer can be less dense because it is hotter). The crust, being lighter, does not participate in the overturn of the layers below it, but is distorted by the overturning material, forming the corona.

Arachnoids: Characterized by a central annulus of fractures or ridges with radial fractures or ridges extending outward from the annulus for several radii (Fig. 9.24). The name *arachnoid* comes from the resemblance of the feature to a multilegged spider on a web. Most annuli have diameters in the range 50–175 km, with only a

(continued)

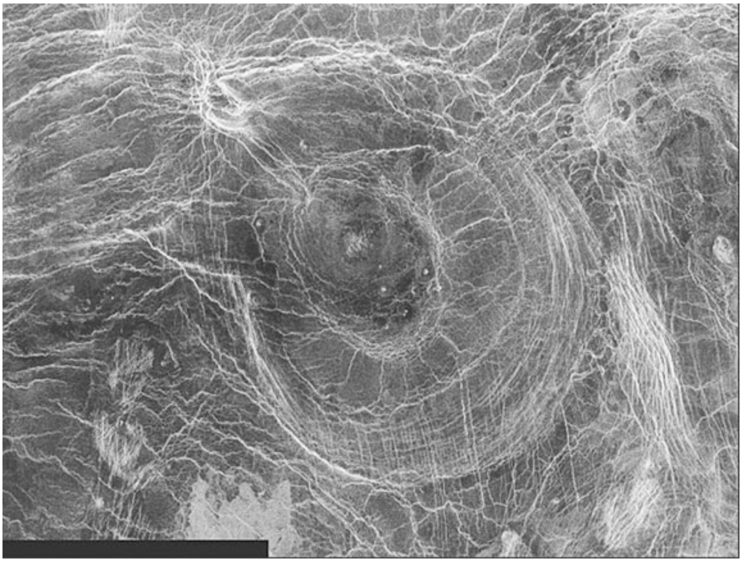


Fig. 9.24 An arachnoid in Fortuna. The radial cracks may be associated with upwelling magma within the dome. *Magellan* NASA/JPL image

(continued)

few over 200 km. The interior of the annulus is a depression of depth 0.5–2 km. Concentric, extensional fractures (where the surface has been stretched) can occur on the depression slope, or concentric ridges where relaxation of the crust into the depression causes compression.

Some differences between arachnoids and coronae are:

Most coronae are >200 km diameter; most arachnoids are <200 km.

Coronae are dominated by concentric features, arachnoids by radial features.

Coronae show extensive volcanism; arachnoids may have a few central shields but lack extensive volcanism.

Corona interiors can be depressed, dome-shaped or plateau-like; arachnoid interiors are always depressed.

Arachnoids are believed to result from diapirism (Fig. 9.23). The magma thins the ductile lower lithosphere (uppermost mantle), allowing the brittle upper lithosphere (crust) to relax downward to form a depression of similar diameter to the diapir. Radial extension at the circumference of the depression causes fracturing of the crust to form the annulus. If the diapir is close to the surface, horizontal dykes radiating from the diapir along the neutral buoyancy level can result in radial fractures. If it is deeper, radial ridges can be caused by tangential stresses as the lithosphere sinks. The lack of extensive volcanism is a consequence of the magma having reached a level of neutral buoyancy at a depth below the surface. See Krassilnikov and Head (2003) for more discussion and references.

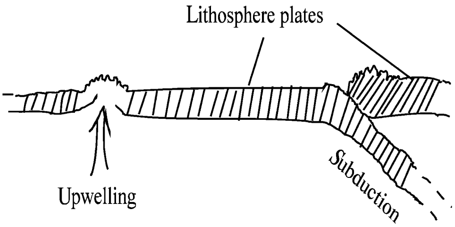


Fig. 9.25 Seafloor spreading and subduction of the lithosphere on the Earth

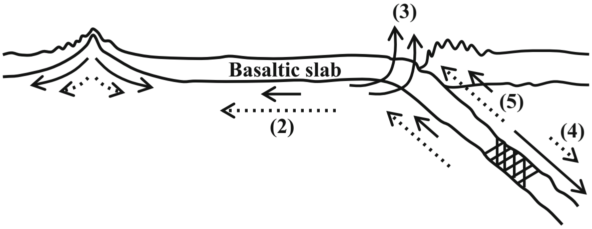


Fig. 9.26 The five forces of tectonics illustrated. *Arrows indicated with numbers* denote forces described under the same number in the text. *Solid arrows*: Earth. *Dotted arrows*: Venus. The longer of the two arrows in each numbered pair is the one believed to be stronger in that pair (e.g., ridge push is stronger on the Earth than on Venus), but arrow lengths are not otherwise related to the magnitude of the force

9.2.3.5 Comparative Plate Tectonics Earth vs. Venus

No evidence has been found for plate tectonics on Venus; in particular, there are

- No interconnected spreading centers (analogous to mid-ocean ridges)
- No subduction trenches or their accompanying chains of volcanoes
- No mountain arcs (like terrestrial island arcs but without oceans) that would indicate plate motion over a mantle hotspot

Plate tectonics requires three conditions:

1. Rigid lithospheric plates
2. Creation of new lithosphere along one boundary of the plate
3. Subduction of lithosphere along another boundary of the plate

On Venus, the lithosphere is not as rigid as on Earth because of the ~450 K higher surface temperature. The lithosphere may therefore deform without being subducted, inhibiting global plate tectonics.

Figure 9.25 illustrates plate tectonics on Earth from the viewpoint of the processes acting on a plate.

Five types of forces are involved in plate tectonics; they are described below and compared to the expected forces on Venus, using the numbers in Fig. 9.26.

(1) **Ridge Push** Forces acting along a mid-ocean ridge push the two plates apart:

(a) The material rising from the mantle physically pushes the plates apart.
(b) The ridge is elevated, so the weight of the plates makes them slide down each side of the ridge, pushing outward on the two plates.

On Earth, the force due to the rising material is larger than the sliding force by about an order of magnitude.

Ridge push may be important in initiating subduction by forcing the far edge of a plate to thrust under an opposing plate.

On Venus, the lithosphere is hotter than on Earth because of the ~450 K higher surface temperature. The implications are that:

The lithosphere is softer and less able to support high elevations.
A "mid-ocean ridge" on Venus would be expected to be only about 40 % as high as on Earth.
It would also be less dense because of thermal expansion of the rock.

Ridge push is therefore expected to be much less effective on Venus than on Earth.

(2) **Resistance to Movement over the Underlying Mantle** For this to be important, the flow speed of the underlying mantle must be *less* than the speed of the plate. (If the mantle were moving faster, it would provide a *driving* force.)

A larger resistance reduces the size of a plate which is capable of sliding as a unit over the mantle.

On Earth, resistance to plate motion is greatly reduced by the partially molten *low velocity zone* (LVZ) below the lithosphere.

Venus appears not to have an LVZ. In this case, the much larger resistance to plate motion would greatly reduce the maximum size of plate possible.

(3) **Resistance to Bending** The lithospheric plate is solid and resists being bent at the subduction zone. The resistance is larger for a thicker plate, but reduced if there is an overlying layer (e.g., ocean water on the Earth).

On Venus, we expect a thinner lithosphere but no oceans, so the bending resistance may be similar to that on the Earth.

(4) **Slab Pull** The lithospheric plate (density ρ_L) rests on and descends into the mantle below it (density ρ_M).

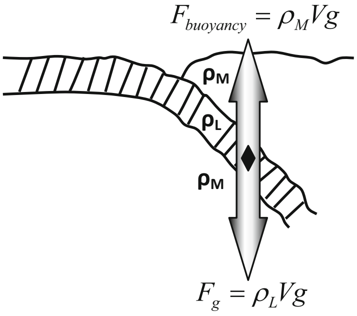
The descending slab feels two forces (Fig. 9.27):

1. The weight downward
2. The upward buoyancy force (equal to the weight of an equal volume of displaced mantle)

The net force (the "*buoyancy*") depends on the difference in density between the mantle and the descending slab, $\rho_M - \rho_L$.

The buoyancy is positive or negative depending on whether this net force is upward or downward:

Fig. 9.27 Slab pull: forces on a descending plate



Case 1: $\rho_\mathrm{L} < \rho_\mathrm{M}$ (positive buoyancy). The slab is less dense than the surrounding material and floats on the mantle.

Case 2: $\rho_\mathrm{L} > \rho_\mathrm{M}$ (negative buoyancy). The slab is denser than the surrounding material and can sink into the mantle.

On Earth, the descending slab is cooler than the mantle it descends into because it has spent 200 million years or so on the surface. As it descends, it is warmed by conduction, but this is a slow process; the slab remains cooler than its surroundings to a considerable depth. Being cooler, it tends to be denser and therefore negatively buoyant. The cooler temperature also causes the basalt-eclogite (at about 50–100 km) and the olivine-spinel (at about 300–400 km) phase changes to occur sooner, increasing the negative buoyancy.

The slab pull force appears to be about ten times larger than the ridge push force, making it the dominant force sustaining subduction once subduction has begun.

On Venus, under a higher temperature,

The lithosphere is less dense due to thermal expansion

The transition from basalt to eclogite is deeper, so the lithosphere remains more buoyant to greater depths

Slab pull may therefore be much less effective on Venus due to the greater buoyancy of the lithosphere.

(5) **Shear Forces on the Descending Slab** These arise from:

Shear between the slab and the opposing plate

Shear between the slab and the underlying mantle

Shear between the slab and the opposing plate is greatly reduced by the presence of water (a major factor on Earth).

On Earth, the shear between the slab and the underlying mantle is apparently zero! This is possibly due to water.

On Venus, there is no water. Shear forces may therefore strongly inhibit subduction.

9.2.4 *Crust*

Three types of crust can be defined for terrestrial planets

1. Primary crust:	Solidification on cooling following accretional heating. Example: the lunar highlands.
2. Secondary crust:	Following partial melting of the mantle; this produces basaltic crust; Examples: lunar maria, Earth's oceanic crust.
3. Tertiary crust:	Following remelting of the secondary crust at the base of a thick crust or on subduction or foundering of the crust into the mantle; this produces granitic crust. Example: Earth's continents.

On Venus, the low crater density, total lack of impact basins >270 km diameter, and obvious resurfacing suggest that no primary crust is left.

9.2.5 *Surface Information from the Soviet Landers*

Much of this material is summarized in Fegley et al. (1997).

Thus far, the only data directly from the surface have come from landers of the U.S.S.R.'s space program.

9.2.5.1 Terrain at the Lander Sites

Venera 8:	*Landed 1972*. Domes and lava flows in plains east of Navka Planitia.
Venera 9:	*Landed 1975*. Steep (15–20°) slope of a hill in Beta Regio, densely covered by ~10-cm-size plate-like rock fragments and some loose soil in the depressions between.
Venera 10, 13, 14:	*Landed 1975, 1982 and 1982, respectively*. Plains, surface dominated by low-standing flat-topped outcrops of bedded rocks with variable amounts of loose soil material in local lows. *Venera 10* was in lowlands near the southeastern edge of Beta Regio. *Venera 14* landed on the flank of a volcano in the south of Navka Planitia.

9.2.5.2 Chemical Composition

This was sampled at seven locations: Veneras 8, 9, 10, 13, 14, and Vegas 1 and 2 sites.

All except *Venera 9* were in plains near the equator. *Venera 9* was on a slope near Rhea Mons (in Beta Regio).

The measured surface temperature was 748 K (475 °C) and the pressure, 90 bars.

The wind speed was less than 1 m/s (3.6 km/h). This wind was strong enough to gradually decrease (over the space of about an hour) the size of a clump of soil which had fallen onto the supporting ring of one of the landers during the landing.

Most of the data indicate that the material around the landers, including bedded rocks, is porous and friable, with densities in the range ~1,400–1,500 kg/m^3, except for *Venera 10* site where the density was around 2,800 kg/m^3.

9.2.5.3 Techniques to Measure Composition

Two techniques were used:

1. *γ-ray spectroscopy*. This was carried out by Veneras 8, 9, and 10, and Vegas 1 and 2 (which landed in/near Rusalka Planitia and carried no TV imagers). It gave the abundances of K, Th, U in the surface layer under the lander.
2. *X-ray fluorescence*. This was carried out by Veneras 13 and 14 and Vega 2. A centimeter-size drill sampled beneath each lander. It gave the abundances of Si, Ti, Al, Fe, Mn, Mg, Ca, K, S, and Cl.

9.2.5.4 Composition at the Soviet Lander Sites

These came from five sites (Veneras 9, 10, 14, Vegas 1 and 2).

The rock was found to be similar to terrestrial *tholeiitic basalt*. Tholeiitic basalt is more silica-rich than other basalts. On Earth it is produced along the mid-ocean ridges and makes up the oceanic crust.

This indicates that the rock represents secondary crust, that is, crust derived directly from material rising from the mantle.

From the individual landers:

Venera 13: This probe landed in the Navka Planitia in the eastern part of the Phoebe Regio. Rocks were found to be similar to terrestrial subalkaline basalts, which are found in rift areas in the Mediterranean region.

Venera 8: Rocks were similar to terrestrial alkaline basalts, and resemble the granitic continental crust on Earth. This was interpreted by some Soviet scientists in the 1980s as indicating ancient crust created by the early differentiation of Venus as it cooled, in a manner similar to Earth (e.g., Surkov 1983). This crust would then have become overlain at a later date by the younger tholeiitic basalts rising from the mantle to form the volcanic regions and the large volcanic plains.

However, the *Magellan* images (Fig. 9.28) show the *Venera* 8 landing site to be in a region of mottled plains with a complex of flows interpreted as basaltic lava flows, related to NW-trending fractures in the plains. Nearby is a pancake dome, 22 × 25 km in extent, made up of a shallow inner depression 10–12 km across surrounded by a raised annulus 5–6 km wide with steep slopes. The inner depression has several shallow, rimless pits. The dome is surrounded with a concentric

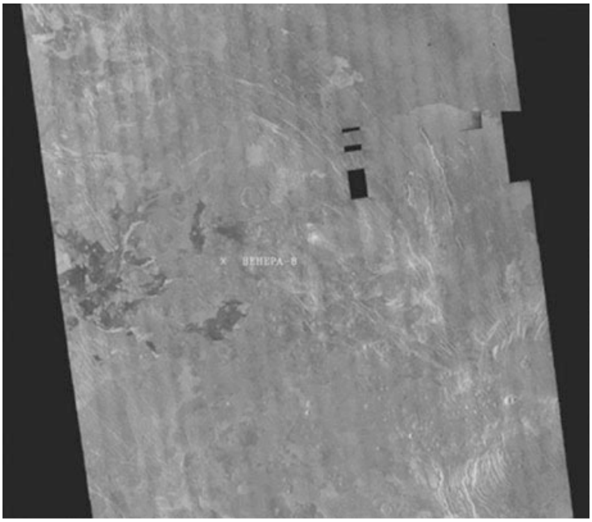


Fig. 9.28 The *Venera 8* landing site in the Navka region on Venus. *Magellan* image, NASA/JPL PIA00460

fracture pattern extending out about 10–15 km. The dome is similar in appearance to rhyolite, dacite, and andesite domes on Earth (rhyolite is the volcanic equivalent of granite). *Venera 8* may thus have measured rocks associated with this dome, rather than ancient crustal rock. This is supported by the young age (<1 Gy) of all surface features seen so far by either *Magellan* or Veneras 15/16.

9.2.6 Pathways for Heat Loss Through Planetary Lithospheres

The three ways in which heat can escape the interior are summarized in Fig. 9.29.

9.2.6.1 Advection

Latin *ad* (to) + *vectum* (carried): The heat is carried to the surface by material transport such as volcanoes, lava flows, and subsurface intrusions of magma.

This is the dominant heat loss mechanism for Io (Jupiter's innermost Galilean satellite; see Chap. 13).

9.2.6.2 Lithospheric Conduction

This is the dominant heat loss mechanism on planets for which plate tectonics, volcanoes and lava flows are not currently important. It is the sole heat loss mechanism for the Moon and Mercury, and the dominant mechanism for Mars. It is also a very important mechanism for Venus.

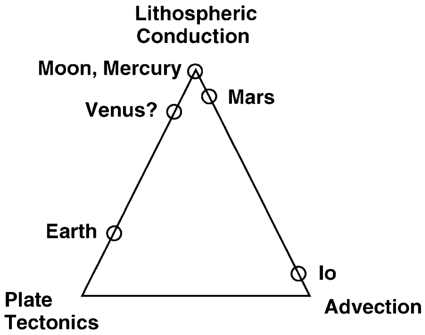


Fig. 9.29 Methods of heat transport to the surface for rocky solar system bodies. The importance of each process for the terrestrial planets and two moons is indicated by the placements of the *open circles*. For the Earth, advection not associated with plate tectonics (e.g., hotspot volcanism) is insignificant compared to the other two processes

9.2.6.3 Plate Tectonics

The mantle is cooled both by advection along the mid-ocean ridges and by heat transfer to the cool, subducting slabs. The Earth is the only known example where this occurs.

9.2.7 Interior Heat Budget of Venus

According to the first law of thermodynamics,[4] the change in internal energy, U, of a system over some length of time, Δt, is equal to the heat input, Q, plus the work input, W, over the same length of time: $\Delta U = Q + W$. The internal energy is the sum of all microscopic kinetic energies (e.g., crystal lattice vibrations) and potential energies within the system. If the system is a planet (e.g., Venus), then the work input, W, can include gravitational potential energy released by contraction as the planet cools, gravitational potential energy released by iron crystals sinking through a fluid core to form a solid, inner core, and heating by tidal friction. (Tidal friction, discussed in Sect. 6.1.1.2, is small for Venus because of its minimal oblateness due to its slow rotation rate, and the lack of a large satellite). The heat input includes the energy, Q_{rd}, released to the planet's interior by the decay of radioactive U, Th and K (positive input) and the energy, Q_{em}, passing outward from the interior to the surface and emitted into space (negative input). Then $\Delta U = Q_{rd} + Q_{em} + W$; or, dividing by the time interval and taking the limit as $\Delta t \rightarrow 0$,

$$dU/dt = P_{rd} + P_{em} + P_{work} \tag{9.6}$$

where P is power (the rate of energy transfer).

[4] For a more general discussion of this principle, see Milone and Wilson 2014, Sect. 10.2.2.

The contribution to P_{em} from advection is determined by the lava temperature and the global rate of lava production; and, from conduction, by the temperatures at the base and surface of the lithosphere and the thermal conductivity. If $|P_{em}| = | P_{rd} + P_{work}|$ then $dU/dt = 0$, in which case the planet is in thermal equilibrium. In this case, the temperature at the base of the lithosphere is constant at some value, T_{equil}. If $|P_{em}| > | P_{rd} + P_{work}|$ (i.e., more energy passes outward through the surface and into space per unit time than is produced in the interior), then $dU/dt < 0$. In this case, the temperature at the base of the lithosphere is greater than T_{equil} (to give the greater outward heat flux) and is decreasing with time. This monotonic decrease in temperature is referred to as *secular cooling*, and arises because the planet is cooling from an earlier, hotter state, e.g., from accretional heating during planetary formation. On Earth, the secular cooling rate is about 100 K per Gyr in the mantle and may contribute up to 50 % of Earth's heat budget.

The rate of heat loss from Venus has not been measured directly, but it can be estimated. One way to do this is to scale from the Earth. The abundances of U, Th, K measured at Venus' surface by the *Venera* and *Vega* landers are similar to the Earth, suggesting a similar rate of radiogenic heat generation per unit mass, while Venus' mass is similar to the Earth ($M_{Venus} = 0.815\ M_{Earth}$), so the accretional heating is expected to have been similar. The Earth's rate of heat loss is 47 TW (Sect. 6.1). Scaling this value by the mass and surface area, the heat flux from Venus' surface would be expected to be ~83 mW/m^2 ± 30 %, the uncertainty being due to possible differences in potassium abundance and differences in secular cooling.

9.2.8 Heat Loss Through the Lithosphere of Venus

With an estimate of the heat flux in hand, one would like to find a model which (a) provides this heat flux through Venus' lithosphere and (b) accounts for the apparently-uniform ~750-million-year age of Venus' surface (Sect. 9.2.3.1). On the Earth, the dominant mechanism of heat transport through the lithosphere is plate tectonics, but no evidence has been found for plate tectonics on Venus. Coronae, which are sites of abundant volcanism, account for ~16 mW/m^2 or less (Smrekar and Stofan 1997) of the total, and the lithosphere is too thick to transport all of the remainder by thermal conduction.

The two most-favored models are,

1. Catastrophic resurfacing (Turcotte et al. 1999). In this model, Venus' lithosphere is currently in a relatively quiet period in which conduction is the dominant mode of heat transport. As noted above, conduction plus advection in coronae cannot transport all of the heat produced in the mantle, so the mantle is increasing in temperature and the lithosphere is gradually thickening. This situation is not sustainable, and at some point the lower lithosphere delaminates in a global subduction event, either accompanied by or followed by large-scale

("catastrophic") volcanism and tectonics. The increased heat loss cools the mantle, initiating the next quiet period. The mean surface age thus varies cyclically, increasing from zero to a final value during each quiet period and being reset to zero by each resurfacing event. It is also possible that only one resurfacing event has occurred, with the preceding regime unknown.

The 2.5-km-thick regional (wrinkle-ridge) plains are generally regarded as the unit responsible for the resurfacing (see Hansen 2007 for references). It is possible that tesserae represent a terrain unit that has survived from before the most recent resurfacing event; see Hansen and López (2011) for discussion.

2. Equilibrium resurfacing (Phillips et al. 1992). Venus in this model is continuously resurfaced by random volcanism over small areas. The mean surface age is then the time required for these small, random events to resurface the planet rather than the time since a single, global event. The mean resurfacing rate equals the surface area of Venus divided by the time over which the resurfacing occurs (~750 million years), or ~0.5 km^2/yr. On the assumption that only 10 % of the magma produced is extruded, with the surface being covered by lava to a depth of ~1 km to cover pre-existing topography, the mean annual magmatic production rate is ~5 km^3/yr, compared to ~20 km^3/yr for the Earth.

9.2.9 Evidence for Current Volcanism on Venus

No direct observation of a volcanic eruption has yet been made on Venus; e.g., no lava flow has been imaged that was not present in earlier images. Nevertheless, at least four lines of evidence suggest that Venus may be volcanically active at the present time:

1. Ultraviolet spectroscopy by the Pioneer Venus Orbiter showed a decline in the abundance of SO_2 from about 100 ppb to about 10 ppb at Venus' cloud tops over the period 1978–1986 (Esposito et al. 1988). This decline was accompanied by a decline in polar haze over the same period. Both of these effects could be explained by ejection of SO_2 into the upper atmosphere by a recent volcanic eruption, followed by its gradual removal by conversion to new, small aerosols of H_2SO_4. As SO_2 decreased in abundance, so would the aerosols, resulting in a decline in the polar haze.
2. Observations by the Visible and Infrared Thermal Imaging Spectrometer (VIRTIS) on the European Space Agency's *Venus Express* spacecraft, which has been orbiting Venus in a highly-elliptical polar orbit since May, 2006, have shown that surface emissivity is significantly greater than average over some lava flows in Imdr, Themis and Dione Regiones (Smrekar et al. 2010). These differences in surface emissivity are caused by compositional differences, either in terms of bulk composition (e.g., basalt *vs.* granite) or the amount of chemical weathering: CO_2 and SO_2 react with minerals (e.g., pyroxene) in basalt to form a thin crust (micrometers thick) that has a lower emissivity than the original minerals. The higher emissivity is consistent with a lack of chemical weathering

and therefore a young age. Also, the youngest craters are believed to be those with dark, parabolic halos formed by wind-blown, fine-grained impact ejecta, with the halos disappearing over time from chemical and/or aeolian weathering. One such dark halo appears to be superposed by one of these lava flows, indicating that the lava flow is younger than the crater.

The ages of these lava flows are believed to be <2.5 million years, and quite likely as little as several hundred to several thousand years. If they are indeed this young, then Venus must still be volcanically active.

3. SO_2 and H_2O are removed from Venus' atmosphere by chemical weathering and loss to space at a sufficiently high rate that continued low-level outgassing by volcanoes within the last 20 million years is required to maintain the present levels of these molecules in Venus' atmosphere (Bullock and Grinspoon 2001). By the same token, Fegley and Prinn (1989) estimate that Venus' clouds would disappear in 1.9 million years without replenishment of SO_2 by volcanism.
4. Bondarenko et al. (2010) find a significant apparent microwave thermal emission excess from a lava flow in *Magellan* images of Bereghinia Planitia obtained in 1993 that suggests still-molten lava below a cooling crust a few meters thick. However, the area does not appear to be otherwise volcanically active; e.g., no volcanoes or coronae.

This concludes our brief summary of the surface science of Venus; as we have already discussed the Earth in earlier chapters, we now turn to the last of the terrestrial planets, Mars.

9.3 Mars

9.3.1 Visibility and General Properties

Ares in ancient Greece, and *Mars* from Roman times, this red planet is associated with the god of war. Its brightness at (mean distance) opposition is −2.01, brighter than all but Venus and Jupiter among the planets that can be seen easily at their brightest. It is the lowest of the classical *superior* planets, i.e., those orbiting beyond the orbit of the Sun in geocentric schemes, and the nearest of the planets exterior to Earth in the heliocentric schemes.

In more recent times it was the center of a controversy about extraterrestrial life. Schiaparelli (~1877) reported the existence of channels *(canali*), visible in telescopic views of Mars in excellent seeing. In the popular English-language press, *canali* became "canals," an interpretation eagerly endorsed by Percival Lowell, who claimed 400 canals could be resolved from Earth. Their widths would have had to be ~50 km or more wide. Lowell thought that the canals channeled water from the melting ice caps to desert-like regions closer to the equator; it was even claimed that a "green wave" swept down from the poles in

Martian springtime. Subsequent work has not confirmed these findings (the "green wave" has been shown to be photometrically gray and space probes have ruled out any such long, linear features), although dedicated Martian ground-based observers will readily concede that in very short moments of superlative seeing, there seem to be many more fine details visible than can be remembered long enough to record in a sketch. These experiences are usually chalked up to psychological or physiological causes, but it would be interesting to see speckle or at least adaptive optics work applied to Mars to see if the lineated features have any basis in integrated images from the planet.

The osculating orbital elements given here are for the date JDN 2456293.5 (January 2, 2013), the epoch. Mars' orbit has a relatively large eccentricity, $e = 0.0933$, and an inclination, $i = 1\overset{\circ}{.}8487$. Its mean distance is 1.5236 au, so its perihelion distance is $a(1 - e) = 1.382$ au, and Mars varies by ±9.3 % in its distance from the Sun. Its ascending node and perihelion longitudes at the epoch were: $\Omega = 49\overset{\circ}{.}5245$ and $\varpi = 336\overset{\circ}{.}0648$. Its sidereal period is 1.88089 y or $686\overset{\mathrm{d}}{.}980$, so that its synodic period is large, in fact the largest of all the planets: $779\overset{\mathrm{d}}{.}94$. The studies of Mars by Brahe and Kepler resulted in the discovery of the elliptical nature of the planetary orbits because of the measurable effects of Mars' large eccentricity.

Mars is a rapid rotator, with a sidereal rotation period, $P_{\mathrm{rot}} = 24^{\mathrm{h}}\ 37^{\mathrm{m}}\ 22\overset{\mathrm{s}}{.}66 = 1\overset{\mathrm{d}}{.}02595675$, compared to Earth's $23^{\mathrm{h}}\ 56^{\mathrm{m}}\ 04\overset{\mathrm{s}}{.}10 = 0\overset{\mathrm{d}}{.}99726963$. Its rotation axis is tilted by $25\overset{\circ}{.}19$ to the axis of the orbit, so that Mars undergoes seasons as does the Earth, but with nearly twice the lengths of Earth's.

It has a radius of 3,397 km, for a mean angular diameter (at mean distance opposition) of $17''.9$. From the semi-major axis of the orbits of Deimos and Phobos, the mass is found from application of Kepler's third law: $M = 6.4191 \times 10^{23}$ kg.

Thus, its compressed mean density is: $<\rho> = 3{,}940$ kg/m^3

This is larger than that of the Moon (3,340 kg/m^3) but less than that of all the other terrestrial planets. Its oblateness or flattening is:

$$\varepsilon = 0.006476$$

corresponding to the second spherical harmonic term of the gravitational potential, $J_2 = \{I_z - I_{xy}]/(Mr^2) = 1.96045 \times 10^{-3}$. This is the *quadrupole moment,* which measures polar flattening—see Sect. 5.3. Compare this value to that for the Earth: $J_2 = 1.082626 \times 10^{-3}$. The J_3 term, which indicates asymmetry between N and S hemispheres is also relatively large: $+36 \times 10^{-6}$ (compared to Earth's -2.533×10^{-6}). The Martian ε and J_2, J_3 values are the largest among the terrestrial planets.

At the epoch 2000.0, the North Celestial Pole (NCP) of Mars was located at $\alpha = 21^{\mathrm{h}}\ 10^{\mathrm{m}}\ 43\overset{\mathrm{s}}{.}54$, $\delta = +52°\ 53'.19$ (Folkner et al. 1997), ~9° NW of Deneb (α Cygni). Due to the precession of the Martian equinoxes, it slowly changes: $d\alpha/dt = -0.1061(7)$ degrees/century, $d\delta/dt = -0.0609(4)$ degree s/century (for both values, the number in parentheses gives the uncertainty in the last decimal).

The mean geometric albedo of Mars is 0.12, and its color indices are: (B-V) = 1.36 and (U-B) = 0.58. These are the reddest colors of all the planets. The general properties of Mars are summarized in Table 9.1.

9.3.2 Martian Geologic Epochs

In terms of age, terrains on Mars have been divided into four different epochs, or periods, based on crater density. An absolute chronology for these epochs depends on the model used for the cratering rate, so the values listed below should be regarded only as approximate; e.g., the Noachian epoch could have begun ~4.1 Gy ago if there was a steady decline in the early bombardment after planetary accretion, or as late as ~3.8 Gy ago if there was a late spike, the suggested Late Heavy Bombardment. (The unit Gy = giga-year = 10^9 years.)

- **Pre-Noachian**. The ~ 500 million year interval between the formation of Mars ~4.6 Gy ago and the age of the oldest surfaces now visible, ~4.1 Gy. Essentially all pre-Noachian surfaces have been covered by lava flows and sediments or erased by impacts and erosion.
- **Noachian**. The oldest visible surfaces and therefore the most densely cratered, dating from ~4.1 to ~3.7 Gy b.p. Most of the southern highland surface (Sect. 9.3.4) is of Noachian age. The Noachian epoch is named after Noachis Terra (“Land of Noah”), an ancient landform in the southern highlands of Mars, west of the Hellas impact basin.
- **Hesperian**. ~3.7 to ~3.0 Gy ago, an interval marked by the emplacement of extensive lava plains. The Hesperian epoch is named after Hesperia Planum, a high plain of Hesperian age in the southern highlands, northeast of the Hellas impact basin.
- **Amazonian**. ~3.0 Gy ago to present. Very low cratering rate, and therefore the least cratered surfaces. The Amazonian epoch is named after Amazonis Planitia, an extremely smooth plains region west of the giant Martian volcano, Olympus Mons. Somewhat ironically, after the Amazonian epoch was named for this region, Amazonis Planitia was discovered to contain some of the smoothest deposits in the solar system, with an age as low as a few tens of millions of years (Fuller and Head 2002). Thus, although of Amazonian age, it is somewhat atypical of the Amazonian epoch.

9.3.3 Principal Types of Surface Features

We now describe the principal surface features on Mars, and then discuss the nature and history the global structures of Mars.

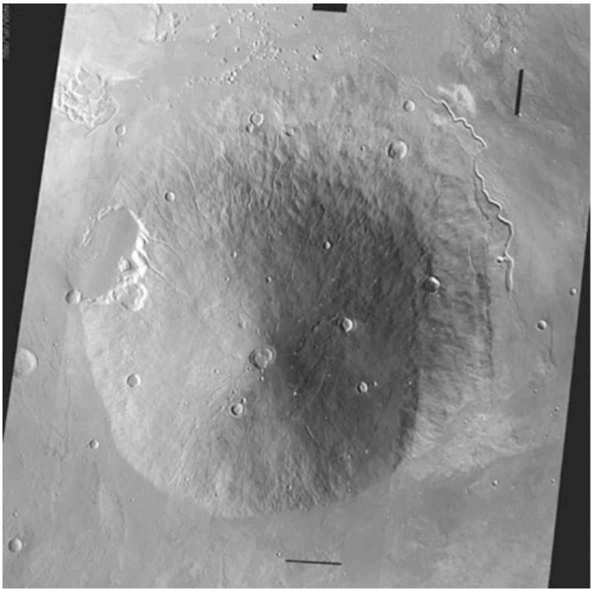


Fig. 9.30 Hecate Tholus, a relatively small volcano; note the many impact craters in the area. A 2001 *Mars Odyssey* THEMIS image, PIA06827. Credits: NASA/JPL/ASU

9.3.3.1 Volcanoes and Related Features

Types of volcanic features seen on Mars include:

Shield volcano:	A volcano with a broad, gently sloping cone; it usually has a shallow caldera at its summit. Examples on Earth: Mauna Loa; Kilauea.
Patera:	A volcano with an even lower profile than a shield volcano, and may be even larger in diameter; it often has a complex caldera at its summit.
Mons:	Literally, “mountain,” but on Mars it applies specifically to the largest shield volcanoes.
Tholus:	A smaller, steeper-sided volcano than a shield volcano.
Pyroclastic volcano:	An explosive volcano, producing ejecta. Earth example: Mt. Vesuvius.

Specific volcanic features include:

Tharsis Bulge. About 4,000 km diameter, rising as high as 10 km above the mean datum.

The three large Tharsis volcanoes, *Arsia Mons, Pavonis Mons,* and *Ascraeus Mons,* lie about 700 km apart along the “summit ridge” of the Tharsis bulge, and there are several smaller volcanoes on the bulge as well. A *tholus* is a volcanic cone (see Fig. 9.30).

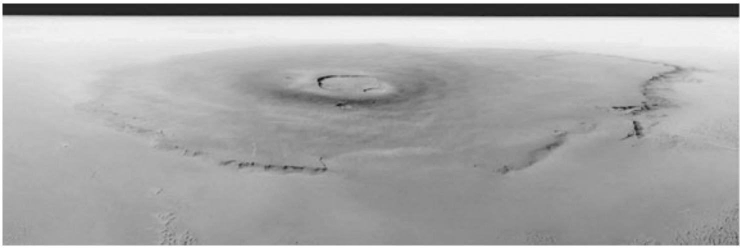


Fig. 9.31 Olympus Mons, at 27 km high, is the highest feature on Mars (dwarfing Mauna Loa and Mt Everest on Earth and Maxwell Montes on Venus). Its caldera is ~80 km across and the entire construct is ~600 km wide. The escarpment itself is 6 km high. Note the clouds around the peak and above the flanks. A *Mars Global Surveyor* mission, Mars Orbital Camera wide field image PIA04737, viewed toward the Martian limb. Release No. MOC2-479, September 10, 2003. Credit: NASA/JPL/Malin Space Science Systems

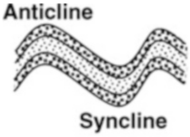
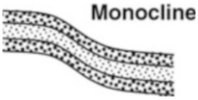
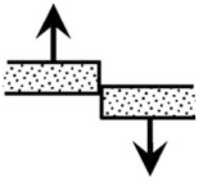
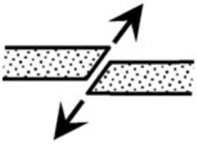
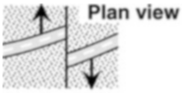
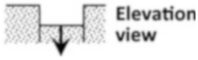
Olympus Mons. The largest volcano on Mars, located NW of the Tharsis bulge, with a base of 600 km diameter and a summit peak of 80 km width and 27 km height. Olympus Mons appears as a small dark spot on older visual maps. At times this spot appears bright because of clouds, inspiring Schiaparelli to name it Nix Olympica, "Snows of Olympus." Schiaparelli noticed that, during dust storms, it was often one of the few features visible, and concluded correctly that it must be an elevated area. The volcano was renamed Olympus Mons after spacecraft revealed its true identity. Compared with the island of Hawaii, the Earth's largest volcanic shield construct, it is about four to six times the diameter (~150 km maximum extent, ~100 km mean extent) and three times the height (~9 km, measured from the seafloor bottom). See Fig. 9.31.

Alba Patera. A huge, ancient shield volcano, in the shape of a shallow dish ("patera"), lying north of the major Tharsis volcanoes and NE of Olympus Mons. It is larger in diameter than Olympus Mons, 1,600 km across, but only about 1/3 the height. Evidently far older than Olympus Mons, it is north of and between Olympus Mons and the Tharsis Ridge volcanoes. Circular cracks around it testify to the great strain on the surrounding material of this volcano, which must have deeply sunk down toward the mantle as it achieved *isostatic equilibrium.*

The Tharsis bulge began to form in the Early Noachian epoch, ~4 billion years ago, although the volcanoes are much younger. The bulge may thus have resulted in some way from the event(s) that created the global dichotomy described in Sect. 9.3.4.

9.3.3.2 Tectonic[5] Features

Figures are elevation views except where noted as a plan view:

Fold:	Deformation of a part of the crust due to cooling and consequent compression.	
Anticline:	In a sinusoidal deformation, the crest of the fold (the “hill”).	
Syncline:	In a sinusoidal deformation, the trough of the fold (the “valley”).	
Monocline:	An incomplete fold, characterized by a graduated change from one level to another.	
Fault (or rift):	A break in the crust (instead of folding).	
Normal fault:	Vertical displacements, in opposite directions, of adjacent pieces of crust.	
Thrust fault (or reverse fault):	One piece of crust rides up over an adjacent piece of crust, which slides under.	
Strike-slip fault:	Horizontal displacements, in opposite directions, of adjacent pieces of crust.	
Graben:	A piece of crust that has dropped between two normal faults.	

Canyon systems. The largest of these is Valles Marineris (45°W to 90°W, near the equator). This is an extensional rift system about 4,000 km long, extending eastward from the middle of the Tharsis bulge. The valley system is up to 500 km wide and 7 km deep (see Fig. 9.32). It includes, among many, the following major canyons:

[5] Tectonic: pertaining to (or caused by or resulting from) structural deformation of the crust.

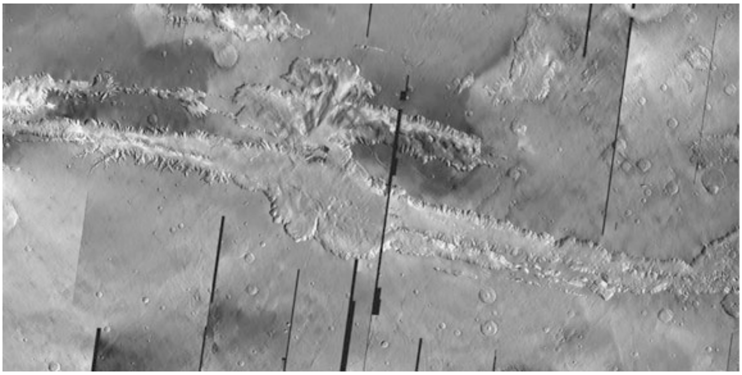


Fig. 9.32 The western portion of Valles Marineris, the major rift system on Mars, extending 4,000 km east–west and as wide as 500 km in places. The widening of the main valley near the *center* of the image is Melas Chasma, with Ius Chasma leading into it from the *left* and Coprates Chasma leading out of it to the *right*. Tithonium Chasma is the narrower valley north of Ius Chasma. Note the chains of collapse pits outside the canyon's south side. The loss of subsurface ice may contribute to the growth of the canyon system. A 2001 *Mars Odyssey* spacecraft mosaic, NASA/JPL/Arizona State University image PIA06926

Ius Chasma (centered ~7° S, ~82° W)
Tithonium Chasma (75 km wide, 4 km deep) (centered ~5° S, ~85° W)
Coprates Chasma (centered ~13° S, ~60° W)

Fossae[6] Extensional fractures (grabens), usually occurring as fracture systems.

9.3.3.3 Depositional, Erosional and Related Features

Fretted terrain. Flat lowlands bordered by steep cliffs and filled with elevated plateaus (mesas). They are due to erosional recession of the highlands.

Chaotic terrain. These are depressed areas characterized by jumbled slabs and blocks. They are likely due to subsurface withdrawal. Examples: *Hydaspis Chaos,* ~100 km wide; *Aureum Chaos* (see Fig. 9.35).

Layered terrain. The earliest known and best examples are near the poles. They are stacked layers, each ~30 m thickness, due mainly to aeolian deposits of sediments. *Mars Global Surveyor* images have also shown layered terrain in basins and craters; these tend to be more frequent at lower latitude sites.

[6] *Fossa* is the Latin word for a ditch or trench. *Fossae* is the plural. See Figs. 9.33 and 9.34 for examples on Mars.

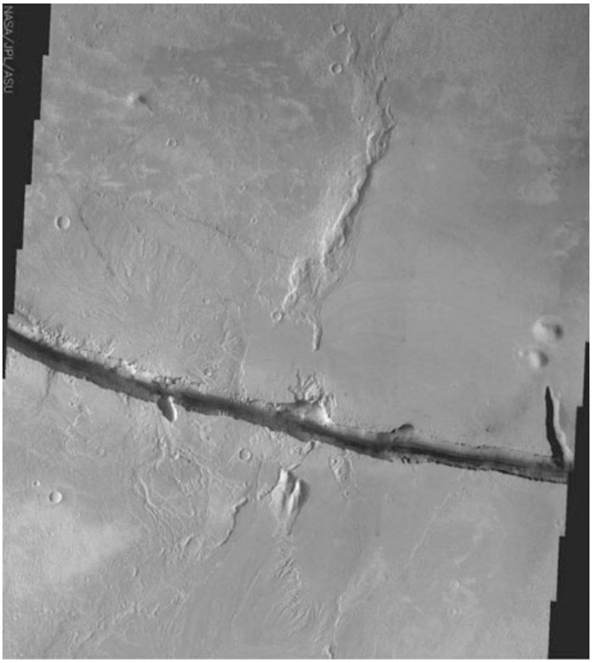
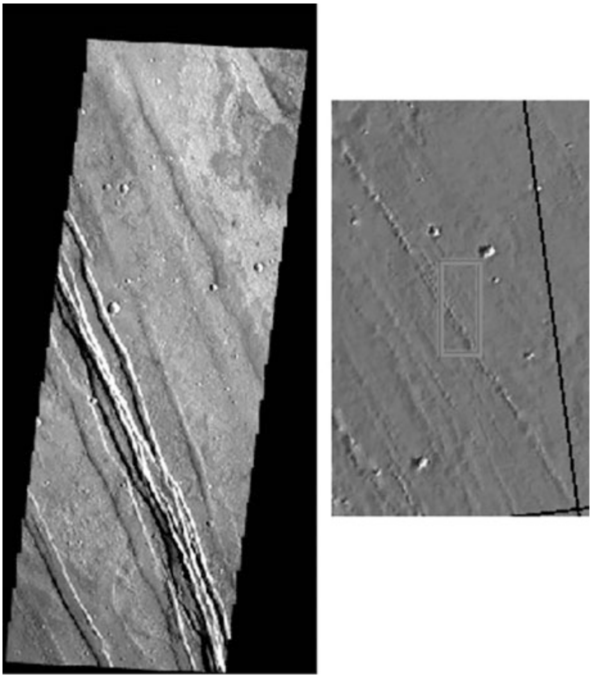


Fig. 9.33 As seen from the *Odyssey* 2001 spacecraft, a graben, part of Cerberus Fossae in a volcanic region of Mars. Note the debris at the bottom of the feature where the surface material has collapsed between two faults. A 2001 *Mars Odyssey* THEMIS VIS instrument image PIA06842. Credits: NASA/JPL/ASU

Fig. 9.34 Multiple parallel faults at 23°S and 259°E in the region between Syria Planum and Claritas Rupes on Mars. A 2001 *Mars Odyssey* THEMIS VIS instrument image, NASA/JPL/ASU image PIA02297, and context image, produced by Arizona State University, Tempe



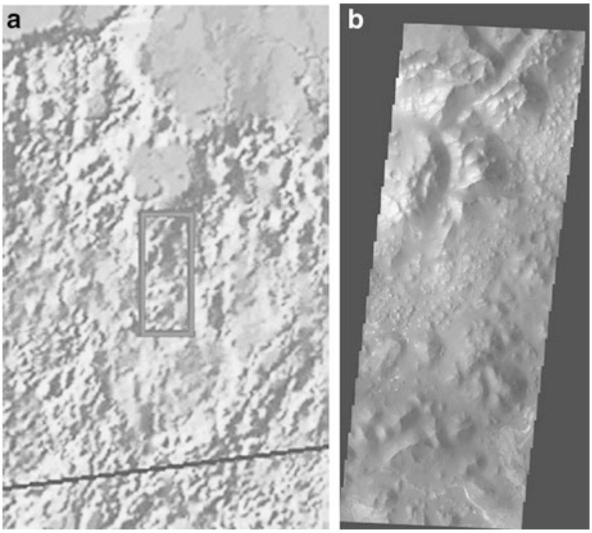


Fig. 9.35 Aureum Chaos, 3.6°S, 333°E. (**a**) Context image, showing the highlands; (**b**) blocks and deposition as seen from the *Mars Odyssey* 2001 spacecraft. THEMIS VIS instrument images, PIA02196. Credits: NASA/JPL/ASU

Etched plains. These are located in polar regions, and are pits and hollows possibly due to saltating particles and sublimation (primarily of CO_2 ice).

Outflow Channels. These large-scale features show the presence of immature tributaries and are widespread. They may be as much as hundreds of kilometers long, starting and stopping abruptly. See Figs. 9.36 and 9.37. Some characteristics (not always present) are:

- Tear-drop shaped islands and bars (as, for example, in the Ares Vallis)
- Meanders (one particularly extensive one closely resembles the Red River on Earth)
- Sometimes braided patterns in the region, resembling post-flood debris deposition
- Sometimes found in chaotic terrain

An example is Maja Vallis in Lunae Planum.

Dendritic channels. These resemble terrestrial stream beds created by surface runoff on sloping terrain. Individual streams merge as the water flows downhill, forming a dendritic (branching) pattern. The dendritic networks on Mars are fairly common in Noachian-aged highland terrain, with lengths up to tens of kilometers. Individual channels have V-shaped cross-sections typical of erosion by a river, and range in size up to ~1 km wide and ~100 m deep (Chapman et al. 2010; Erkeling et al. 2010). Based on terrestrial analogues, sustained fluvial erosion for times up to ~10,000 years is required for a dendritic network to mature to the form seen.

The dendritic channels west of Echus Chasma in the Kasei Valles area north of Valles Marineris were carved in Middle Hesperian deposits, ~3.4 Gy b.p. If the water resulted from precipitation rather than subsurface sources or melting of

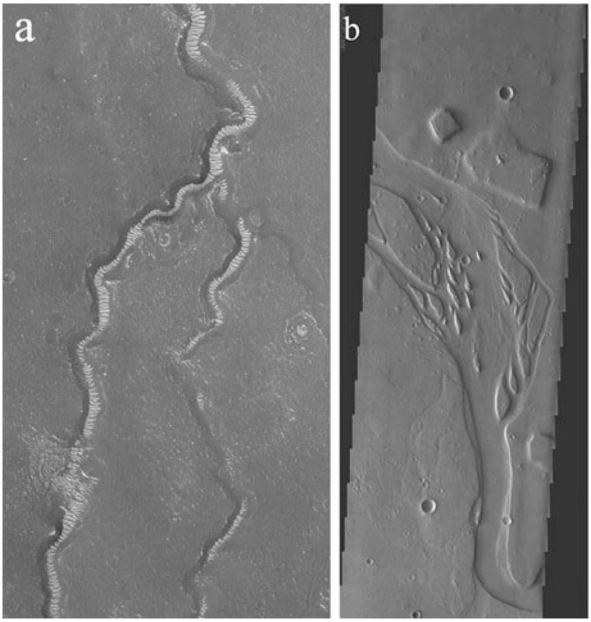


Fig. 9.36 (**a**) Apsus Valley channels, with meanders and sand dunes. MGS Mars Orbital Camera image, PIA05992. Credits: NASA/JPL/Malin Space Science Systems. (**b**) SE of Elysium Mons, with tear-drop islands. 2001 *Mars Odyssey* THEMIS images, PIA04586. Credits: NASA/JPL/Arizona State University (ASU)

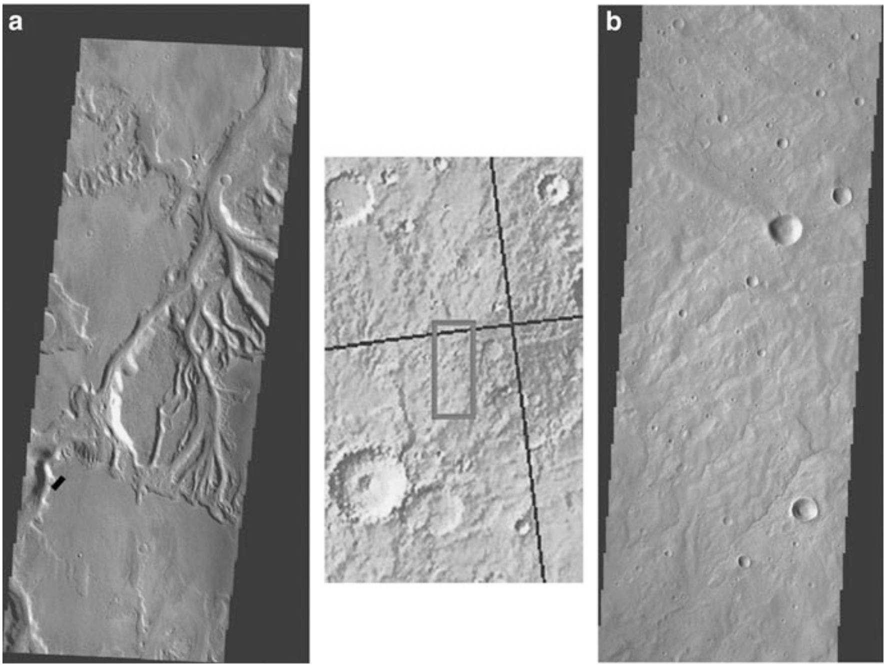


Fig. 9.37 (**a**) The Sabis Vallis. Note the tributaries. Image PIA03663. (**b**) Channels joining to form Sabis Vallis. Image PIA03664, the context for which is shown in the *middle*. Credits: NASA/JPL/ASU. 2001 *Mars Odyssey* THEMIS VIS images produced by Arizona State University (ASU)

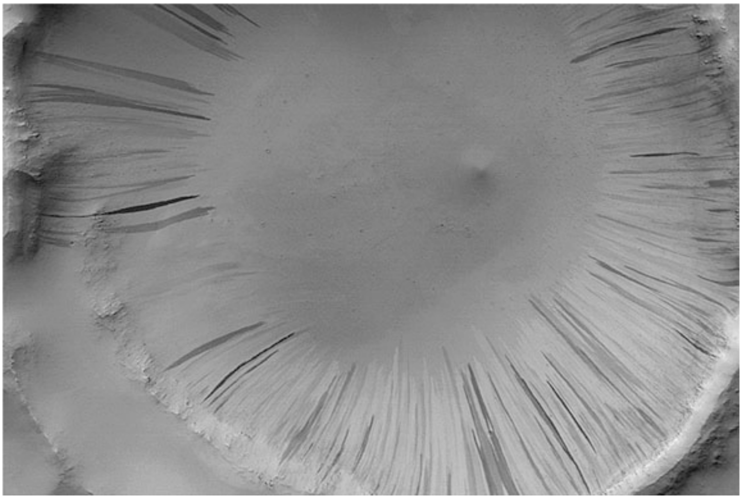


Fig. 9.38 Recurrent slope linneae in an approximately 4-km diameter crater near 13.0 N 319.8 W, where either fluid material originated from the subsurface at or below the top of the crater wall and flowed downhill into the crater, or solid material has fallen and, saltating, precipitated avalanches. MOC narrow-angle image R05-00278. Credits: NASA/JPL/Malin Space Science Systems

glaciers, then it may have been caused by local climate change from Tharsis volcanic activity (Chapman et al. 2010). If so, then rainfall conditions must have recurred into the Hesperian, at least locally.

Gullies or arroyos. These resemble drainage channels, and are found on crater walls and elsewhere. On some crater walls, they appear to emerge from a particular layer; these regions do not appear very ancient, for the most part. Outbursts of both CO_2 and H_2O have been suggested as possible causes. The features are found more frequently at higher latitude sites, which tend to be well below the freezing point of water, thus favoring landslides created by outbursting pockets of CO_2. On the other hand, the ice content of the surface layers is expected to be more deeply buried nearer the equatorial regions, where a higher rate of sublimation would long ago have exhausted the near-surface ice.

Gullies may be straight, sinuous, or irregular, and with or without debris aprons. Among the straight type are *recurring slope linneae (RSL)*. These are small-scale seasonal flow features up to ~5 m wide and several tens of meters long that are found on crater walls and other steep slopes (<20°) in the southern highlands between ~30° and ~50° S latitude (Ojha et al. 2012). They appear in the Martian spring, gradually extend downslope until late summer, and fade in fall and winter (Fig. 9.38). Thus, they indicate possible liquid-water processes occurring on Mars today, such as the melting and flow of brine (Chevrier and Rivera-Valentin 2012). Most (there can be thousands along a crater wall) prefer equator-facing slopes where temperatures reach 250–300 K, and all confirmed RSL descend from a bedrock exposure, often in association with a small gulley. However, the conditions

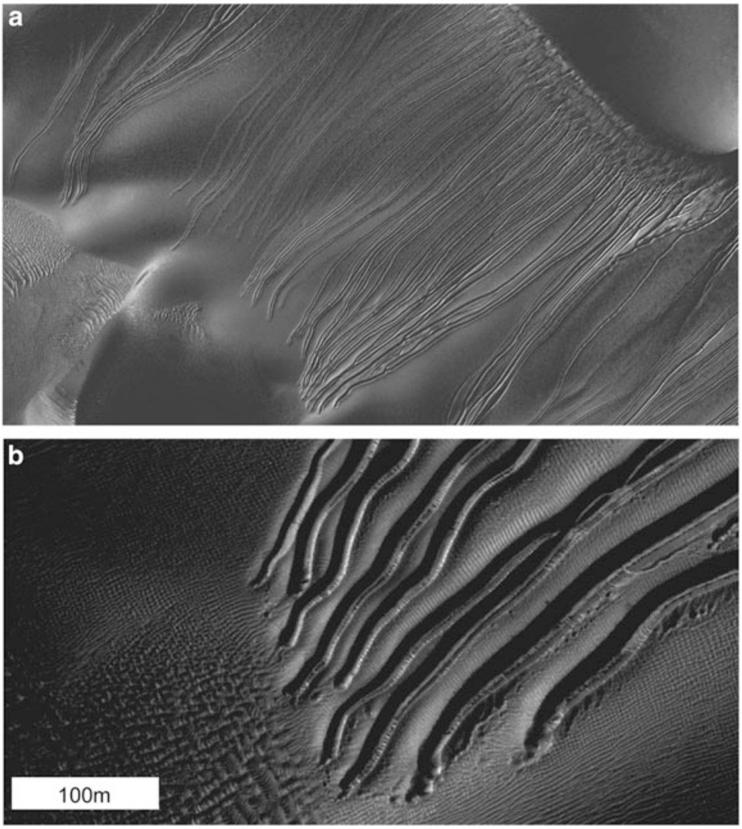


Fig. 9.39 (**a**) Linear gullies in Russell Crater at 54°.25 S and 12°.92 E, on Mars. Some gullies extend as far as 2 km down the side of the large dune seen here. (**b**) An enlarged view of the lower ends of the gullies at the *bottom center* of the image in (**a**). The raised banks with little width variation and minimal or absent debris aprons at the terminations are characteristic of this type of gully, which may be caused by blocks of CO_2 coasting downhill on a cushion of subliming gas. PIA 17260, a portion of HiRise image PSP_001440_1255 obtained on November 16, 2006, from NASA's Mars Reconnaissance Orbiter. Courtesy of NASA/JPL-Caltech/U of Arizona

listed here cannot be the only ones determining the occurrence of RSL, because >200 features with similar characteristics have been examined that show no signs of RSL (Ojha et al. 2012).

Gullies can also form in completely-sandy conditions, most commonly on dune slopes (Fig. 9.39) but also on sandy crater walls. These *linear gullies* are characterized by nearly-uniform widths of a few to 10 m, lengths from a few hundred meters to ~2.5 km and depths usually from <1 m to 2 m, but can be >3 m in places (Diniega et al. 2013 and references therein). Usually a gulley starts at the dune brink with a small alcove and/or converging small grooves, and terminates with a pit or a

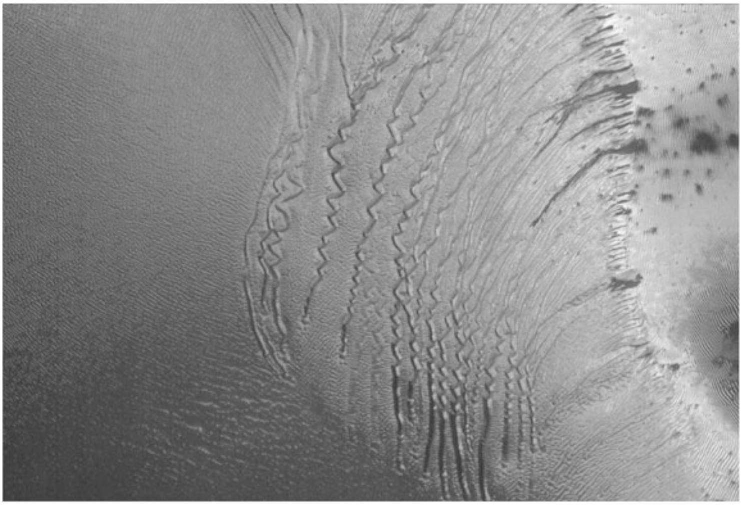


Fig. 9.40 Extremely sinuous linear gullies on a dune in Kaiser Crater at 47°.24 S and 19°.48 E, on Mars. A terminal pit is visible at the end of most of the gullies. A portion of HiRise image PSP_010749_1325 obtained on November 11, 2008, from NASA's Mars Reconnaissance Orbiter. Credits: NASA/JPL/University of Arizona

chain of pits, or sometimes a series of divergent small grooves, each with a terminal pit. Paths are often sinuous, sometimes very strongly so (Fig. 9.40). New gulleys have been seen to form and existing grooves to elongate at the start of each spring, indicating that these are active features. These linear gullies are distinguished from *alcove-channel-apron* features that terminate in cones or aprons characteristic of debris flows, without meandering.

A possible origin for the linear gullies seen in Figs. 9.39 and 9.40 is in blocks of CO_2 (dry ice) that break from the winter accumulation of CO_2 at the top of the dune as the temperature rises in spring (Diniega et al. 2013). It is suggested that these blocks slide down the slope on a cushion of subliming gas, creating gullies similar to the width of the block as they descend. CO_2 ice can accumulate to a depth of a few tens of centimeters at mid-latitudes to 1–2 m in the polar regions.

Araneiform terrain: Martian "spiders" (Fig. 9.41). There is some debate about how these features form. One suggestion is that they are created by pockets of expanding CO_2 beneath a thick but transparent layer of frozen CO_2 ("dry ice"). Kieffer et al. (2006) present evidence that in Martian winter at the southern polar cap, the ice is basically spotless, but, with spring sunrise, dark spots begin to appear, followed by fan structures, and then spider-like features. The visual transparency of the CO_2 allows the dusty substrate to absorb sunlight, and become warmer; this warms the substrate, causing the dry ice above it to sublime. They interpret the spots as jets of dust-laden CO_2 erupting through the overlying dry ice, the fans as wind-

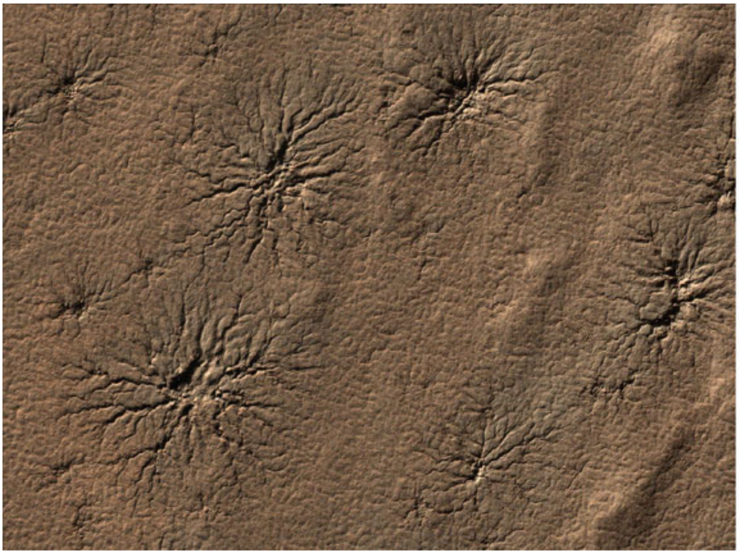


Fig. 9.41 Isolated araneiform ("spider-like") features on Mars, thought to have been excavated in a dusty substrate of the south polar region on Mars by evaporating carbon dioxide gas overlain by a mantle of clear "dry ice." The strip is ~1.2 km across and the depth of the grooves of the spider's legs is typically1–2 m. The altitude of the Sun was 15° in this late springtime image, from which all traces of CO_2 ice have vanished. A portion of image ESP_014413_0930 taken on August 23, 2009, at 87.0° S and 86.5° E by the High-Resolution Imaging Science Experiment (HiRISE) camera on NASA's Mars Reconnaissance Orbiter (MRO). NASA image PIA 12249. Courtesy of NASA/JPL-Caltech/University of Arizona

blown deposition, and the spiders as channels left behind in the dusty substrate by gas flowing toward the eruption site. This hypothesis was vigorously challenged by Langevin et al. (2006), who found only weak features of CO_2 in reflection spectra. However, Kieffer et al. (2006) attribute this to contamination by dust and an overlying coating of water ice. Langevin et al. (2006) argue that the required level of dust contamination to explain the spectra would reduce the sunlight by a factor of 3, making it difficult to create the sub-ice bubbles in the polar region in the first place. They also question why no such phenomena are seen elsewhere on Mars where the clear-ice CO_2 signature is stronger.

Dust deposits. *Barchan dunes* (crescent-shaped with apex facing the wind), *transverse dunes* (rows of overlapping crescents, apices all facing the wind), and *longitudinal dunes* (more or less parallel, linear, carved by shifts in wind direction) are seen in many areas, including floors of wide craters. Sometimes dusty regions can be identified by contrast to dark streaks downwind of isolated dome or mesa-like structures, which serve to block aeolian (windborne) deposits. Figure 9.42 illustrates a variety of dunes.

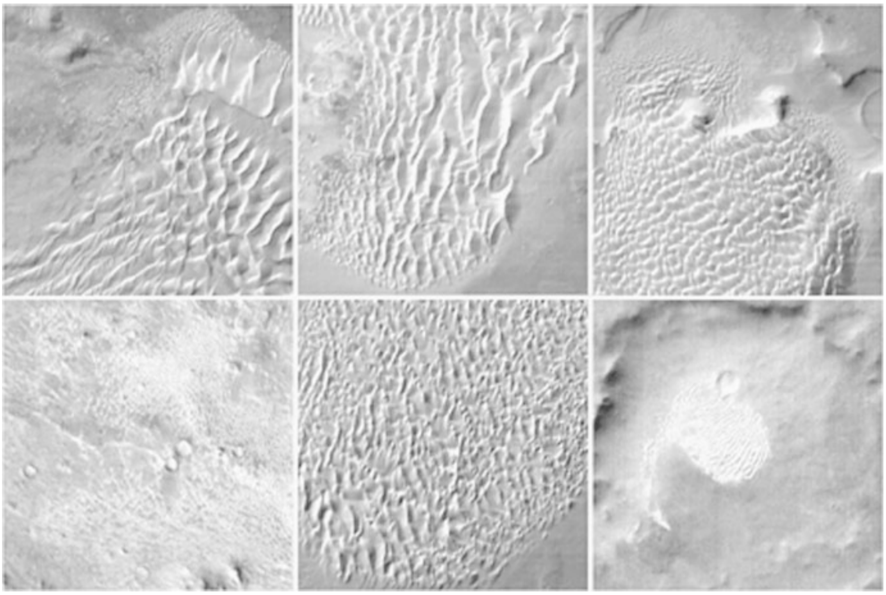


Fig. 9.42 Infrared (12.6 μm) images of dunes from various areas of Mars. NASA/JPL 2001 *Mars Odyssey* THEMIS VIS image PIA03740 produced by Arizona State University

9.3.3.4 Impact Basins

There are 19 known impact basins over 250 km diameter; the following five are among the largest. For comparison, the Moon's Mare Imbrium has a diameter of 1,300 km.

Hellas	(centered ~40° S, ~295° W)	2,300 km dia
Isidis	(centered ~15° N, ~270° W)	1,900 km dia
Argyre	(centered ~50° S, ~40° W)	1,200 km dia
Prometheus	(centered ~83° S, ~270° W)	850 km dia
Chryse	(centered ~20° N, ~45° W)	800 km dia

The Argyre basin is prominent in Fig. 9.43.

An analysis of topological data led Andrews-Hanna et al. (2008) to conclude that a vast region in the north was in fact a major impact site. The size of the resulting basin is 10,600 × 8,500 km, centered on 67° N, 208° E (see the accompanying map, Fig. 9.44). This would make the *Borealis basin* not only the largest impact feature on Mars, but in the entire solar system. The authors state that an impactor nearly 2,000 km across is required to create it. Such an impact could possibly resolve the Martian dichotomy issue which we discuss next.

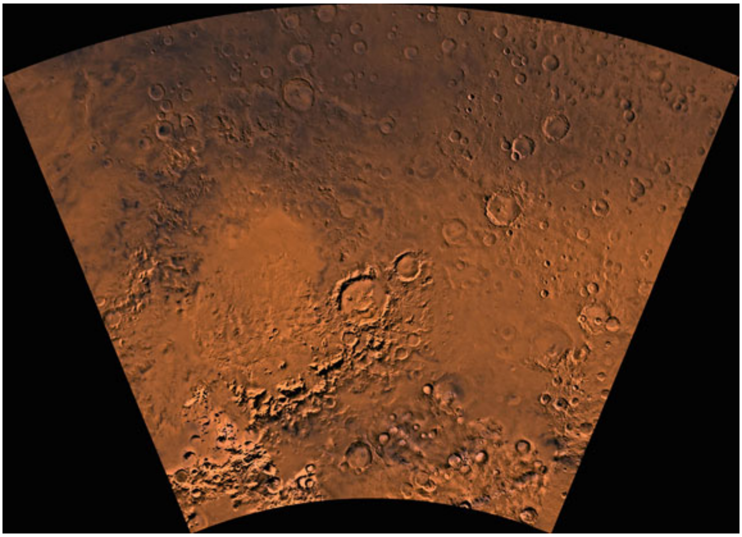


Fig. 9.43 The Argyre region of Mars. North is up. The large crater just below center is Galle, 215 km diameter. The nearly-circular, ~800-km-diameter expanse of light-colored plains west of Galle is Argyre Planitia, which covers the interior of the large Argyre impact basin. The rim of the basin is marked by an approximately-circular border of rugged mountain blocks that includes the Charitum Montes to the south of Argyre Planitia and the Nereidum Montes to the north and northwest. Latitude range 65º to 30º S, longitude range 0 to 60º W. Viking-1 Orbiter image mosaic produced by the U.S. Geological Survey, image PIA00186, MC-26 quadrangle. Credit: NASA/JPL/USGS

9.3.4 The Global Dichotomy: Structure and Origin

The J_3 value given in Table 9.1 and Sect. 9.3.1, above, indicates that Mars has a large global dichotomy. This is reflected in the physical terrain:

- The southern hemisphere is primarily old, heavily cratered terrain of early Noachian age, lying 2–3 km above the "mean datum" for Mars.
- The northern hemisphere is primarily younger, lightly cratered terrain (the Vastitas Borealis Formation) of Middle to Late Hesperian age, lying below the mean datum. The density of cratering is about 2–3 times that on the lunar maria.

An excellent review of this dichotomy and models for its formation is provided by Watters et al. (2007). Further discussion and references may be found there.

The boundary between the southern highlands and the northern lowlands is partially obscured by the Tharsis bulge (Sect. 9.3.3.1), but if we neglect this bulge then the boundary is approximately a great circle making an angle of about 30° with the Martian equator. Over much of its length, the boundary is marked by a relatively steep scarp with an elevation difference of ~2.5–3.5 km, although in the vicinity of the heavily-cratered highland area of western Arabia Terra it is <1 km.

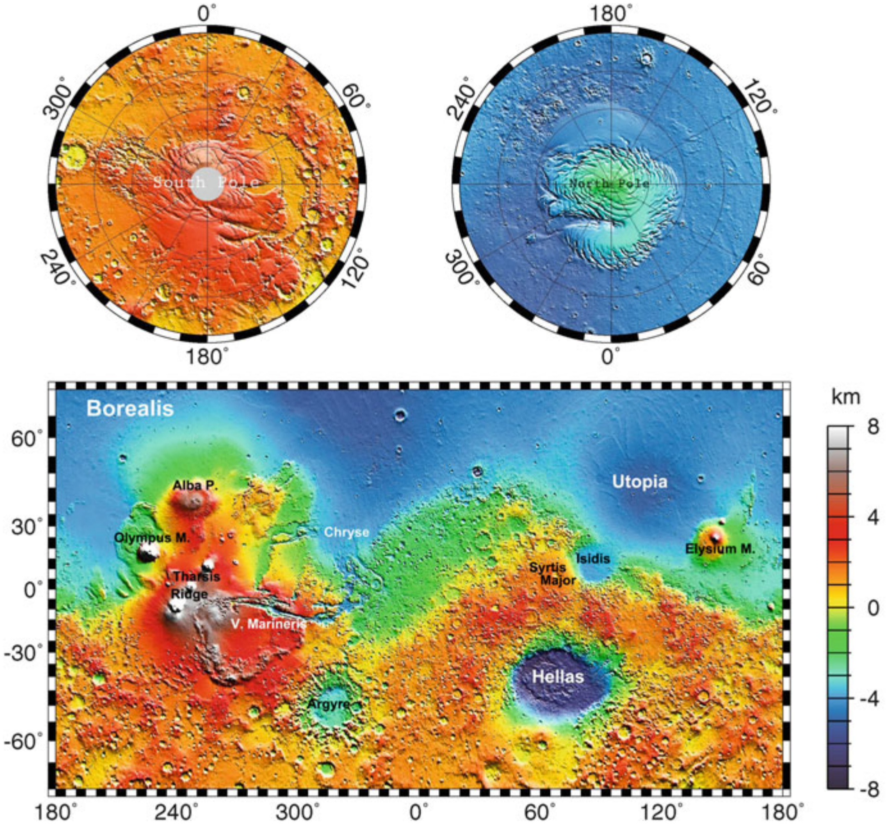


Fig. 9.44 Global topographical maps of Mars. *Top*: South (left) and North (right) polar regions in stereographic projection. *Bottom*: Mercator projection to ± 70^0 latitude. Prominent basins, mountains, and the great Valles Marineris and Tharsis ridge are labeled. Note the strong dichotomy between the north and south hemispheres. With greater topographical range, an elliptical form for the proposed Borealis basin becomes visible (Andrews-Hanna et al. 2008). The scale at *right* indicates the elevation above or below the aeropotential surface that corresponds to the 6.1 hPa atmospheric pressure isobar (Read and Lewis 2004). Abbreviations: M. = Mons, P. = Patera; V. = Valles. Data from the Mars Orbiter Laser Altimeter (MOLA) instrument on the Mars Global Surveyor. Goddard Space Flight Center, NASA image PIA 02031. Credit: NASA/JPL/GSFS. Added annotations by EFM

Away from the boundary, the elevation difference between the southern highlands and northern lowlands is everywhere >2.5 km, and >6 km in places. This difference is comparable to that between the continents and ocean floor of the Earth (generally 4–6 km).

The origin of the dichotomy is still uncertain. Endogenic models, i.e., those that involve processes within the planet, that have been proposed include (1) removal of the lower crust by mantle convection, with the lowlands forming by isostatic readjustment, (2) an early period of plate tectonics, and (3) mantle overturn following solidification of a magma ocean. Exogenic models include (1) excavation of the northern lowlands by one or more large impacts, in which case the southern highlands represent the original surface, and (2) a very large impact in the southern hemisphere, with impact melt forming the southern highlands by megadoming (Reese et al. 2011). The northern lowlands then represent the original surface.

High-precision topographic data from the *Mars Orbiter Laser Altimeter* (*MOLA*) on NASA's *Mars Global Surveyor* and subsurface sounding by the *Mars Advanced Radar for Subsurface and Ionospheric Sounding* (*MARSIS*) instrument on ESA's *Mars Express* have revealed buried impact basins in both the highlands and lowlands that are not apparent in visible-light photography. Crater size-frequency distributions for both the visible surface and the buried crust show that the buried lowland crust is actually Early Noachian in age and older than the surface age of the highlands, although possibly younger than the buried highland crust (Frey 2006). Thus, rather than being a late feature as suggested by the low crater density on the lowland surface, the global dichotomy is almost as old as Mars itself.

The MOLA data also show subtle, ridge-like features in the *Vastitas Borealis Formation* (*VBF*) of the lowlands, with too low a profile to have shown up in *Viking* orbiter photographic imagery. These features are interpreted to be the surface expressions of wrinkle ridges on lava flood plains that are buried at least 100 m below the VBF surface. They are in fact continuous with exposed Early Hesperian aged wrinkle-ridged plains, up to several kilometers thick, that make up ~10 % of the Martian surface (e.g., Hesperia Planum, after which the time period was named, and Syrtis Major Planum). Including the buried portion, the wrinkle-ridge plains cover >30 % of the surface of Mars, and indicate a globally-significant period of volcanic resurfacing in the Early Hesperian.

The deposits covering the highland crust consist of ejecta from the large, basin-forming impacts in both the highlands and the lowlands. These impacts occurred during the Noachian epoch, hence the Noachian age of the present highland surface. The Vastitas Borealis Formation in the lowlands (overlying the wrinkle-ridged plains, which in turn overlie the lowland crust) appears to be sedimentary, but its origin is still being debated (Sect. 9.3.5).

The topographic dichotomy is also reflected in the thickness of the Martian crust: ~60 km mean crustal thickness in the highlands with a maximum of ~90 km, compared to ~30 km mean thickness in the lowlands with a minimum of ~5 km. The boundary marked by crustal thickness generally follows that of topography except in the vicinity of western Arabia Terra, where the crustal thickness is more representative of the lowlands. It is possible, therefore, that western Arabia Terra is actually exposed, highly-cratered northern lowland crust that escaped volcanic resurfacing.

The great age of the lowland crust implies only a geologically-short time interval between the formation of Mars and the creation of the dichotomy. This may argue against an endogenic origin because of the time required for internal processes to act. On the other hand, the dichotomy boundary is not circular, either as defined by topography or by crustal thickness, as might be expected for a single, major impact, and if the dichotomy formed by multiple impacts, it is difficult to see why these would occur in only one hemisphere. Thus, the origin of the dichotomy is still unclear.

9.3.5 Evidence of Climate Change on the Martian Surface

In the late nineteenth and earlier twentieth century, speculation was rife that Mars was a dynamic place where water was scarce but skillfully managed by an engineering civilization. Following the first spacecraft views provided by the early *Mariner* missions, Mars was viewed as a dead world, one characterized by little atmosphere and waterless deserts that were dominated by enormous numbers of impact craters. Since then, the view has gradually shifted to that of a planet currently caught in a deep ice age, but which may once have possessed a much wetter and warmer climate.

The dendritic channels discussed in Sect. 9.3.3.3 appear to have formed from rainfall and surface runoff during the Noachian epoch. The present Martian climate does not support rain, so these ancient dendritic networks suggest that the Noachian climate was very different from the present. However, there is still considerable debate over exactly what these conditions were and what caused them.

A difficulty in modeling the Noachian climate is that, from stellar evolution computations, the Sun had only ~70 % of its present luminosity early in its life. Thus, it is not enough to increase the atmospheric mass and add enough greenhouse gases to raise the present Martian atmosphere above 273 K, one also has to do this with Mars receiving only 70 % of the solar power that it does now. This is often referred to as the "faint young sun" problem.

Possibilities for the Noachian climate include,

(a) *Continuously warm and wet*. Impacts and/or volcanic activity (e.g., Tharsis) may have released enough greenhouse gases to sustain a warm climate through the Noachian epoch, with a hydrological cycle supporting rainfall, rivers and lakes, and possibly a northern lowland ocean. Conditions such as this would be favorable to the development of life. It has proven difficult, however, to overcome the faint young sun problem (Toon et al. 2010).

(b) *Episodically warm and wet*. Mars may always have been predominantly cold and dry, with warm, wet conditions existing only intermittently; e.g., as a result of the large, basin-forming impacts in the Noachian epoch (Segura et al. 2002, 2008; see also the review by Toon et al. 2010). Water released and heated by a large impact could form a steam-laden atmosphere capable of producing a total of ~600 m of rainfall, and the greenhouse effect from the water vapor might last long enough to allow a hydrogical cycle to develop, further extending the period of rainfall. Lakes and flowing rivers would have existed only during the warm period after an impact. The development of life would not be as favored as in point a, but might survive below the surface during the cold periods.

If any sediments were deposited by a Noachian ocean, they are now hidden below the early-Hesperian wrinkle-ridged volcanic plains and the middle- to late-Hesperian Vastitas Borealis formation (VBF). The large outflow channels that empty from the highlands into the lowlands are of a similar age to the VBF, and

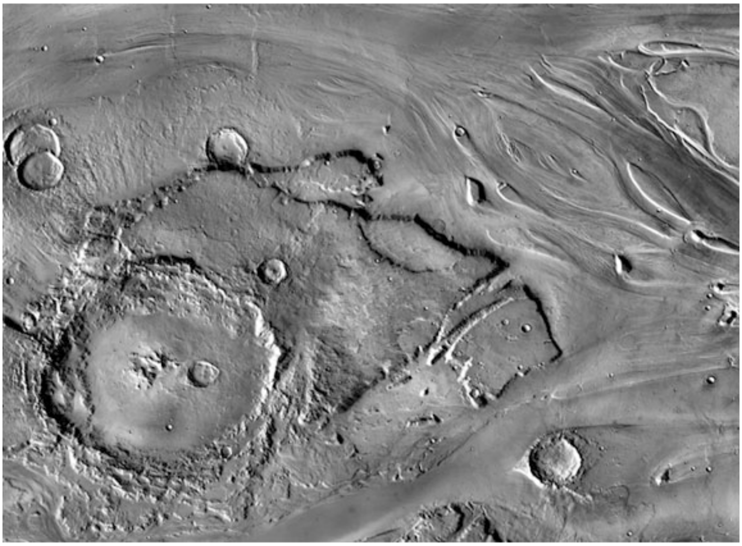


Fig. 9.45 A *Mars Odyssey* mosaic of the north-eastern end of Kasei Valles, near where the valles enter Chryse Planitia. The large crater on the *left side* of the image is Sharonov (dia. 95 km). North is up. Massive floods in Kasei Valles have passed both north and south of Sharonov, creating flow features and teardrop-shaped islands (*top right*). Courtesy of NASA/JPL-Caltech/Arizona State University

may be the source for the sedimentary deposits making up the VBF (Kreslavsky and Head 2002, and references therein). These channels were excavated by massive floods (Fig. 9.45), possibly as a result of the melting and release of large, subsurface aquifers in the highlands.

Other indications of ancient, wetter conditions on Mars include:

1. *Extensive areas with clay minerals in Noachian-aged southern highland terrains*. e.g., Mawrth Vallis, NW of Tharsis, where the clay deposits cover an area $>10^6$ km^2 and can be a few hundred meters thick. They are rare to absent in Hesperian- and Amazonian-aged lowland terrain. Martian clay minerals are primarily Fe- or Mg-rich, e.g., Montmorillonite and nontronite, but in some locations include Al-rich clays such as kaolinite.

   On Earth, clays most commonly form from chemical weathering of igneous rocks by weakly acidic water, most commonly carbonic acid (H_2CO_3) formed by water reacting with atmospheric carbon dioxide:

$$H_2O + CO_2 \rightarrow H_2CO_3 \rightarrow H^+ + HCO_3{}^-$$

   where $HCO_3{}^-$ is the negative bicarbonate ion. Then, e.g., orthoclase ($KAlSi_3O_8$) can undergo hydrolysis with acidic water to form kaolinite ($Al_4Si_4O_{10}(OH)_8$) and quartz (SiO_2):

$$4KAlSi_3O_8 + 4H^+ + 2H_2O \rightarrow 4K^+ + Al_4Si_4O_{10}(OH)_8 + 8SiO_2$$

The presence of abundant clay minerals on Mars has therefore been regarded as evidence for abundant free water in Mars' past, and a climate conducive to life. Clays can also be produced by subsurface hydrothermal activity, but this would not be expected to produce the quantity found. However, Meunier et al. (2012) have suggested that the Fe- and Mg-rich clays are formed during solidification of basaltic magmas. As the magma cools, olivine and pyroxene crystallize first, forming a matrix with pores. If the magma contains water, then these pores can be filled with a hydrosaline solution that precipitates clay minerals as it cools. In this case, the clay deposits would give no information about climate. One aim of the *Curiosity* rover, currently investigating Gale crater where there are thick layers of clay, is to help to resolve these and other issues regarding clay deposits.

2. *In situ analysis of sulfates and hematite at Meridiani Planum*. The exposed bedrock in this area is sandstone, and was studied in detail by NASA's *Opportunity* rover in a 7-m-thick outcrop in Endurance crater (Squyres et al. 2006). Crater counts date the bedrock to the Noachian epoch (Lane et al. 2003). Three layers of sandstone are visible in the outcrop, consisting of sulfate-rich silicate grains. The sulfates include jarosite, predominantly in the form hydronium-jarosite, $(H_3O)(Fe^{3+})_3(SO_4)_2(OH)_4$. Jarosite precipitates only from acidic solutions, and therefore indicates significant amounts of free water, the acidic component perhaps coming from sulfur-bearing volcanic gases reacting with water vapor. Alternatively, if a sulfide such as pyrrhotite ($Fe_{(1-x)}S$, where $x = 0$ to 0.2 and arises from iron vacancies in the crystal lattice) is present, its oxidation in water produces SO_4 while providing the acidity by releasing H^+ ions that attack the silicate surfaces (Chevrier et al. 2010).

   The sandstone also contains significant quantities of hematite (Fe_2O_3), mostly as a mineral in the rock matrix but also occurring as roughly-spherical concretions up to ~5 mm diameter that are distributed quite uniformly through the sandstone. Very fine-grained hematite (<10 μm diameter) appears red because of how it scatters light (these grains in Martian dust give Mars its red colour), but larger grains of crystalline hematite are gray. Gray can look bluish against a reddish background, so these concretions are commonly referred to as "blueberries," and large numbers can be found on Meridiani Planum where they have eroded out of the sandstone (Figs. 9.46 and 9.47). The occurrence of hematite in the matrix and as concretions is consistent with precipitation of the hematite in standing or slowly-moving groundwater.

   The data from *Opportunity* indicate that the original surface layer in Meridiani Planum was olivine basalt (Squyres et al. 2006). This was chemically weathered by acidic water (weak sulfuric acid). Evaporation of the water in a dry climate left an accumulation of sulfate salts and fine-grained silicates, like a *playa* (salt flats) on Earth. While the climate remained dry, erosion of the playa produced sulfate-rich silicate grains that were transported by wind, creating a field of

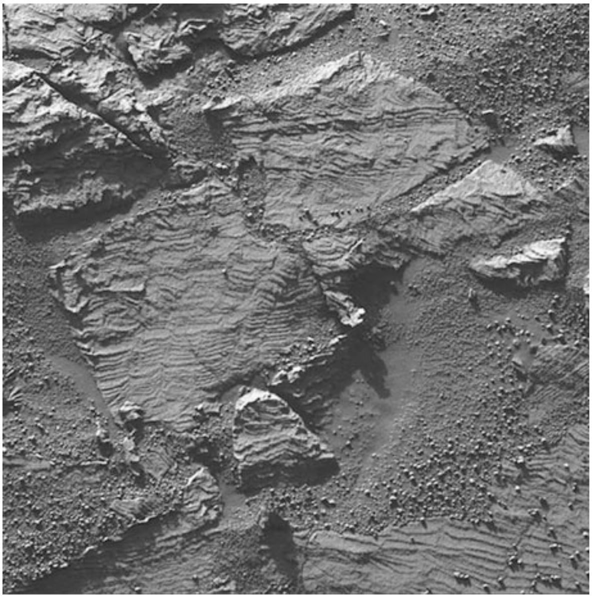


Fig. 9.46 From the Mars Exploration Rover (MER) *Opportunity* in Meridiani Planum: "blueberries," concretions formed in porous rock, and "festoons" (sinuous markings), thought to arise from ripples in an aqueous environment. Courtesy of NASA/JPL-Caltech/Cornell. PIA03279

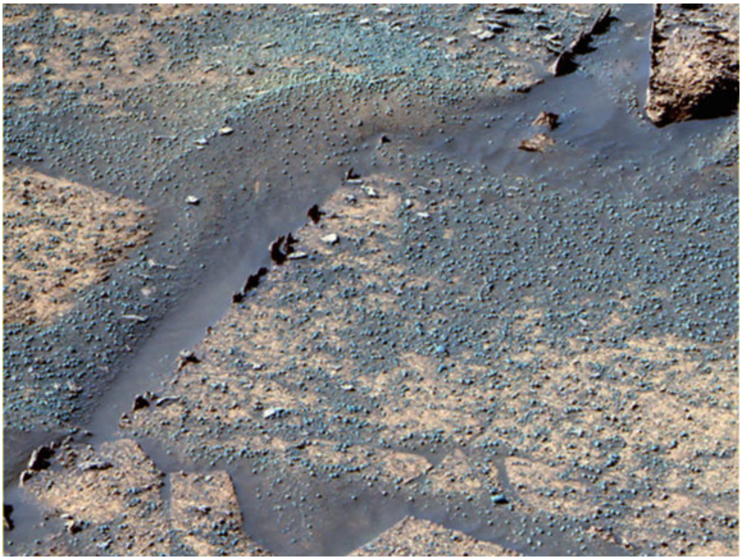


Fig. 9.47 The Mars Exploration Rover *Opportunity* found evidence of a water environment in Endurance Crater in the form of flow channels and "blueberries," shown here. Credits: NASA/JPL/Cornell. Image PIA06692

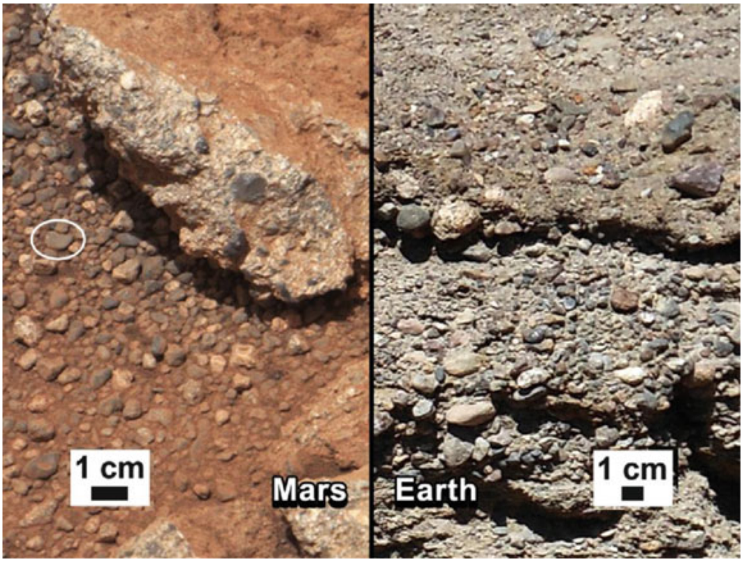


Fig. 9.48 *Left*: A sedimentary conglomerate outcrop in Gale crater on Mars, consisting of rock fragments cemented together, as imaged by *Curiosity* rover. Individual clasts falling from the conglomerate have formed the loose gravel pile on the left. The *circle* highlights a piece of gravel ~1 cm in length. The rounded nature of the clasts shows that they were abraided by collisions during transport in a fluid. The clasts are too large to have been carried by wind, so the outcrop is an ancient streambed. *Right*: A similar deposit on Earth, formed by transport in a stream. Courtesy of NASA/JPL-Caltech/Malin Space Science Systems and PSI

dunes. Subsequently, episodic occurrences of subsurface groundwater (perhaps from occasional flooding?) submerged the sand to varying degrees, at times rising high enough to produce pools of water between the dunes and for at least one period covering the dunes entirely. (The lower levels show characteristics of Aeolian deposit, but the uppermost part of the upper layer shows ripples and cross-laminations typical of water moving over sand.) Interaction of the water with the sulfates cemented the sand and produced the hematite. The geological record indicates, however, that the climate by this point was arid and acidic most of the time.

3. *Ancient streambed gravels*. In 2012 the *Curiosity* rover revealed, within the Gale crater where it landed, evidence of a stream bed replete with rounded pebbles (see Fig. 9.48). The stream bed is in an alluvial fan below a channel that descends from the rim of Gale crater. The pebbles, up to 4 cm across, have been rounded by colliding with other pebbles while being transported by fluid flow over a long period of time. They are too large to have been transported by wind, and are interpreted as having been carried by water with a depth between ankle- and hip-deep flowing at a speed of ~1 m/s (Williams et al. 2013).

Layering has been cited as evidence for a widespread and long-lasting water environment, but polar strata (Fig. 9.49) can be understood as the seasonal interplay

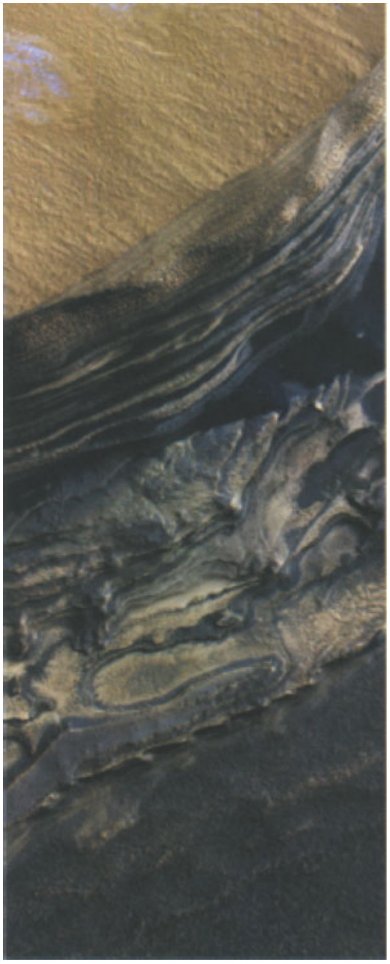


Fig. 9.49 A sub-frame of an image from the Mars Reconnaisance Orbiter High Resolution Imaging Science Experiment (HiRISE), in false color to delineate reflectance from different materials. This view of a scarp of Chasma Boreale, near the north pole of Mars shows layered deposits overlying darker material. Note the interweaving of bright, ice-laden layers with darker sand layers. Image PIA09097, courtesy of NASA/JPL-CalTech/ University of Arizona

of volatile ices (on Mars, water vapor and carbon dioxide) as they sublime and recondense, and the widespread and voluminous dust that coats it during the seasonal global sandstorms. Similarly, the layering seen in canyon walls (Fig. 9.50) could have been laid down in a liquid medium, but aside from that in the strata studied by the Rovers, the large-scaled rock layers could well be products of aeolean deposition.

Extensive subsurface ice existed in the past even at low latitudes, as can be seen from the lobate ejecta around a number of craters (Fig. 9.51). The lobate patterns resemble mud flows, and probably formed by impact melting of permafrost layers.

There is also strong evidence for subsurface ice outside the polar regions at the present time. Neutron spectrometer data from *Mars Odyssey* show abundant subsurface hydrogen poleward of ~50° N and S latitude, which from thermal models is

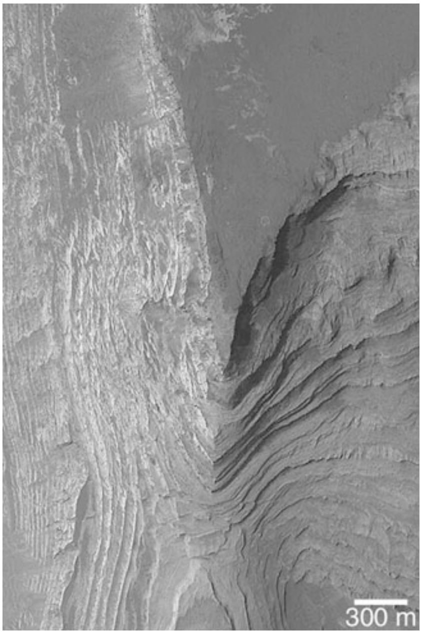


Fig. 9.50 Layering at a precipice at Terby Crater, north of Hellas, near 27.6S, 286 W. *Mars Global Surveyor* Mars Orbiter Camera image PIA04582. Credits: NASA/JPL/Malin Space Science Systems

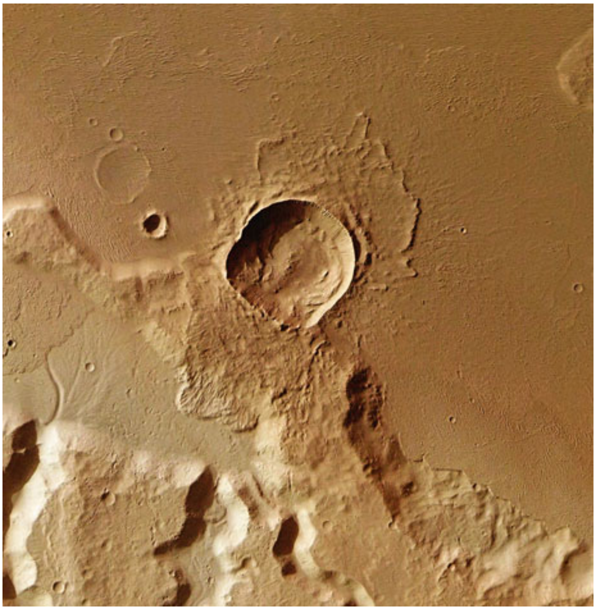


Fig. 9.51 A lobate ejecta pattern can be seen around this ~9-km-diameter crater located near 5° S and 213° E in Medusa Fossae, south-east of Olympus Mons. Taken with the High-Resolution Stereo Camera aboard ESA's *Mars Express* spacecraft. Credit: Image: FUB; Data: ESA/DLR/FUB; courtesy, Prof. S. van Gasselt, Institut für Geologische Wissenschaften, Freie Universität Berlin. Our use was made possible through the kindness and concern of Ms. Stefanie Pott, Sekretariat Planetelogie, FUB

best explained by ice-cemented soil overlain by ice-free soil (Mellon et al. 2004). The depth of the ice below the surface varies from a few millimeters to a few meters, with a few centimeters being most common.

Byrne et al. (2009) observed ice directly by locating five fresh, small craters (4–12 m diameter) in *Mars Reconnaissance Orbiter* high-resolution images in the range 43°–56° N latitude, that showed bright deposits on their floors. "Before" and "after" images show that the crater-forming impact occurred between the two images, and subsequent images showed fading of the bright deposits over several months at a rate consistent with sublimation of water ice. This composition was confirmed in one crater, where the spatial extent of the exposed material was great enough to obtain a spectrum showing water-ice absorption bands. Crater depths ranged from 0.5 to 2.5 m, so the original depth to the ice was less than these values. Numerical models gave depths to the ice in the range 12–70 cm.

A strong indication of localized subsurface water ice on poleward-facing slopes at south latitudes as low as 25° comes from *Mars Express* and *Mars Reconnaissance Orbiter* IR data (Vincendon et al. 2010). The instruments measure solar radiation reflected from the surface, and so do not detect the ice directly; however, the condensation of CO_2 ice on the surface is controlled by the amount of heat escaping from below, and this is greater if there is water ice within a few meters of the surface than if there is not. (Ice or ice-cemented soil has a much greater thermal inertia than loose soil, so as the temperature drops to the CO_2 condensation temperature, water ice remains warmer longer than loose soil.) From numerical energy-balance computations, the distribution of winter CO_2 ice at low latitudes is best fit by a subsurface model with ice at a depth below the surface varying with latitude from 6 cm at 45° S to 13 cm at 35° S.

Another indication of present-day subsurface ice is the appearance of many possibly water-fed gullies in the walls of craters, suggesting that some kind of interior warming may act from time to time to sublime or melt ice below the surface, which can then flow through porous strata and cause the observed features. From thermal profile models, Mars is estimated to have a heat flux output of 30 $\mathrm{mW/m^2}$ (Carr 1999). This helps to explain why gullies tend to be seen emerging from particular strata in features located at higher latitude sites, where the subsurface ice is found closer to the surface, than from sites near the equator. Figure 9.52 demonstrates gullies at several sites.

9.3.6 Mars' Formation and Interior Structure

Numerical simulations indicate that the formation of a terrestrial planet can be divided into three stages (see Dauphas and Pourmand (2011) for discussion and references):

- Formation of planetesimals 10–100 km diameter from the solar nebula.
- Collisions of planetesimals over a period of a few million years to form planetary embryos. This stage begins with a period of runaway growth in which gravitational focusing factors increase with mass; i.e., the more massive an object becomes, the faster it grows. However, above a certain mass, the embryos

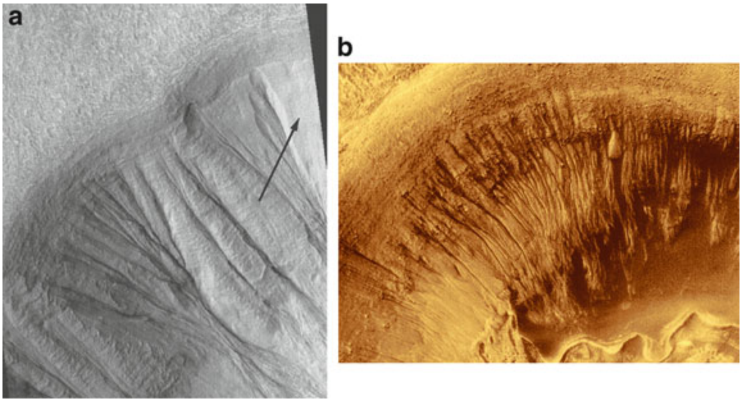


Fig. 9.52 Mid-latitude crater wall gullies. *Mars Global Surveyor* Mars Orbiter camera images. (**a**) Image M09-2875 (PIA04409) of a site 33°S and 93°E showing a 2.8 × 4.5 km view. The arrow indicates a possible snowpack source for the outflows. Credits: NASA/JPL/Malin Space Science Systems/Philip Christensen. (**b**) Image MOC2-242A PIA02824/PIA01039. The north wall of a 7-km diameter crater within the 287-km wide Newton crater at 41°S, 160°W. The gullies are characterized by alcoves due to subsidence, at the source, the outflow channels (of water or CO_2), and the apron of debris at the base. Similar characteristics are seen on such features on Earth. Credits: NASA/JPL/Malin Space Science Systems

begin to gravitationally clear their orbits of planetesimals; consequently, the more massive they become, the slower they grow (a period known as oligarchic growth). This oligarchic growth limits the final masses of planetary embryos to a range similar to the masses of the Moon and Mars.

- Chaotic growth by collisions of planetary embryos over several tens of millions of years to form planets. In the case of the Earth, the final such collision (the one that formed the Moon) has been dated to >50 million years after the start of planetesimal formation (Sect. 8.7).

The decay of ^{182}Hf to ^{182}W (half-life ~9 million years) can be used to date core formation and therefore the maximum formation time of a planet, because of the siderophile and lithophile natures of tungsten and hafnium, respectively (Sect. 8.7). The decay of ^{146}Sm to ^{142}Nb (half-life 103 million years) is also useful. Dauphas and Pourmand (2011), from a precise isotopic analysis of Martian meteorites (Sects. 15.1.3.2 and 15.3.2 in Milone & Wilson 2014), find a formation time for Mars of only ~4 million years. This short a formation time suggests that Mars reached only the end of the second of the three stages described above: i.e., Mars managed to avoid collisions with other planetary embryos and thus failed to grow further (or be annihilated!).

A consequence of Mars' short growth time is that there would still be significant amounts of the short-lived radioactive isotope ^{26}Al (half-life 7.17 × 10^5 yr) in the accreting material. The energy released by this decay would be enough to give Mars a deep magma ocean, and the impacting planetesimals would also be molten.

Two seismometers have been placed on the Martian surface, attached to the legs of the *Viking* landers, of which only one returned data. A single candidate seismic event was found in 500 Martian days of observations. For the future, the design for the *InSight* mission (***In****terior Exploration using* ***S****eismic* ***I****nvestigations,* ***G****eodesy and* ***H****eat* ***T****ransport*), currently projected to launch in 2016, includes a highly-sensitive seismometer. In the absence of a seismic array, information about the Martian interior has been obtained by *in situ* analysis of Martian rocks and soil by robotic landers and rovers, laboratory analyses of Martian meteorites, remote sensing by orbiting spacecraft, and measurement of the moments of inertia of the planet and the potential Love number k_2 (Sect. 5.3). The latter quantities provide information about the mass distribution in the Martian interior and whether the core is rotationally decoupled from the mantle; i.e., whether the core is at least partially molten.

Orbiting spacecraft can observe the entire Martian surface, but care must be taken to allow for weathering. Acidic weathering increases the silica content of the rock surface by preferentially leaching olivine from the surface layer; this (for example) makes basalt look like andesite when viewed remotely with *Mars Global Surveyor's* thermal emission spectrograph (McSween et al. 2009). The Mars Exploration Rovers *Spirit* in Gusev Crater (2004–2010), *Opportunity* on Meridiani Planum (2004–present), and *Curiosity* in Gale Crater (2012–present), can investigate only relatively small areas of the surface, but have the advantage that they can brush away accumulated Martian dust and also grind through or, in the case of *Curiosity*, evaporate (with a laser) a weathered surface to expose unweathered rock.

9.3.6.1 The Martian Crust

Several aspects of the Martian crust were discussed in Sect. 9.3.4. Here, we add a few more points to this discussion.

The dominant rock type on Mars is tholeiitic basalt, similar to terrestrial basalts of mid-ocean ridges, seafloor bedrock, and oceanic islands above mantle hotspots (e.g., Hawaii) (McSween et al. 2009). Martian basalt has a much higher FeO content than on Earth: ~16–18 wt% compared to ~10 wt% on the Earth.

Basalt is a volcanic rock, so on the assumption that the dominant surface rock type reflects the composition of the crust, the Martian crust was formed by lava flows from magma source regions in the mantle after the magma ocean had solidified. Primary crust like that of the lunar highlands, which formed by flotation of anorthositic plagioclase feldspar during solidification of the lunar magma ocean, appears to be absent on Mars. Two possible reasons for this are, from Elkins-Tanton et al. (2005),

- Mars has a much greater amount of water than the Moon, and water in a magma ocean inhibits formation of plagioclase.

- Above a certain pressure, plagioclase reacts with olivine to form pyroxene and an aluminous phase such as spinel or garnet (Gardner and Robins 1974), which would sink in a magma ocean. Because of the greater gravitational acceleration on Mars than the Moon, this pressure would occur much closer to the surface in a Martian magma ocean.

9.3.6.2 The Martian Mantle and Core

From tracking data for *Mars Reconnaissance Orbiter*, *Mars Global Surveyor* and *Mars Odyssey*, Konopliv et al. (2011) obtain a potential Love number $k_2 = 0.159 \pm 0.009$ and a polar moment of inertia $C/(MR_e^2) = 0.3644 \pm 0.0005$, where M and R_e are the mass and equatorial radius, respectively, of Mars. If the core were solid, k_2 would be <0.08 (Rivoldini et al. 2011), so the value of k_2 confirms that the core (or at least the outer core) is fluid. Konopliv et al. (2011) obtain a core radius in the range 1,630–1,830 km without comment about a solid, inner core, whereas Rivoldini et al. (2011) with a somewhat different analysis using the same k_2 value obtain 1,729–1,859 km and a completely fluid core. A completely fluid core is consistent with the present absence of a global magnetic field because (once a planet has cooled from its initial post-accretion state) a solid, inner core is believed to be necessary for the vigorous convection required by a core dynamo; see Sect. 9.1.4.

In the Earth, the lower mantle consists primarily of perovskite, obtained by phase transitions from pyroxene and olivine (Sect. 7.4). The presence or absence of a perovskite layer is an important factor in the convective behavior of the mantle. The lower gravitational field strength of Mars (0.38 that of the Earth at the surfaces of the two planets) means that pressure increases less rapidly with depth in the Martian mantle than in the Earth's mantle. Consequently, a given phase transition occurs at a much greater depth in the Martian mantle than on Earth; e.g., the olivine-spinel transition (~400 km depth on the Earth) would be at >1,000 km depth on Mars, the precise depth depending on temperature and mantle iron content. Whether the spinel-perovskite transition (~660 km depth on Earth) exists at all in the Martian mantle depends on the size of the core; if the core is too large, the required pressure will not be reached in the mantle. Rivoldini et al. (2011) find that, with their predicted core size and for reasonable temperature and iron abundance ranges near the core-mantle boundary, a thin perovskite layer is possible at the base of the mantle if the temperature is near the hotter end of the range, but not otherwise.

The presence of a perovskite layer, with its endothermic phase boundary, would tend to suppress convection in the Martian mantle (Sect. 7.8). The suppression is stronger when there is a shorter distance between convective plumes (Zuber 2001); i.e., when there are more plumes. Thus, one or two plumes are more likely than six or seven. Numerical simulations in fact suggest that the presence of a perovskite layer would result in a single, stationary and long-lived mantle plume. This has

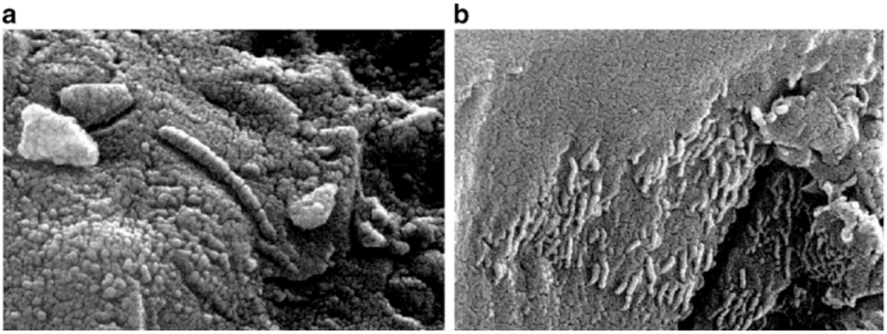


Fig. 9.53 Possible morphological evidence for the presence of microfossils (or parts or fragments thereof) in segmented carbonate deposits in the Martian meteorite ALH 84001. NASA/AMLAMP images (**a**) PIA00288 and (**b**) PIA00284

been suggested as a model for the Tharsis rise, which formed in the Noachian epoch and remained volcanically active into the geologically-recent past (Zuber 2001 and references therein).

We conclude this chapter with a brief discussion of the possibilities for life on Mars, either past or present.

9.3.7 The Search for Life on Mars

The announcement at a NASA press conference in 1996 of the possible discovery of evidence for past life on Mars was electrifying. McKay et al. (1996) cited mineralogical and morphological evidence for their conclusions about a meteorite from Mars recovered from Antarctica, ALH 84001 (described in Sect. 15.3 of Milone and Wilson 2014). Current evidence includes the presence of magnetite, possibly biogenically-produced, and polycyclic aromatic hydrocarbons (PAHS). However, a number of objections have been raised: possible non-biogenic origin for the 10–200 nm-scale carbonate (Treiman et al. 2002; Treiman 2003) and magnetite (Golden et al. 2001) structures; the segmented forms (Fig. 9.53) were but artifacts of the imaging process (Bradley et al. 1997); such small organisms could not contain the structures needed to metabolize; and the terrestrial contamination of the material in the sense that a significant fraction of the carbon present is ^{14}C, dispersed from atomic bomb tests of the 1950s (Jull et al. 1997); and other, mineralogical arguments. Although vigorously defended (e.g., by McKay et al. 1997; McKay et al. 2002; Thomas-Keprta et al. 2002), the present day community, on the whole, has not accepted the claim.

The availability of liquid water is important when considering possibilities for life. Independently of the interpretations of the structures in ALH 84001, the carbonates in which they occur provide a window into the environmental conditions in which these carbonates formed. ALH 84001 is ~90 wt% orthopyroxene, 2 %

chromite, ~2 % shock-produced feldspathic-glass, ~1 % carbonates, and traces of other materials. The ^{176}Lu-^{176}Hf radiometric age of the igneous components is 4.1 Gy (Lapen et al. 2010); i.e., the time at which the rock crystallized from magma (see Milone and Wilson 2014, Sect. 15.5 for a discussion of radiometric ages). From ^{87}Rb-^{87}Sr dating, the carbonates precipitated in the period 3.9–4.0 Gy b.p. (Borg et al. 1999). Both the rock and the carbonates thus formed during the Noachian epoch (Sect. 9.3.2). The ratios of various stable isotopes, e.g., ^{18}O/^{16}O and ^{13}C/^{12}C, depend on the environmental conditions in which the carbonates formed; e.g., temperature, and the availability of reservoirs such as the atmosphere. Isotope ratios in the carbonates indicate precipitation at an approximately-constant temperature of 18 $\pm$ 4°C (Halevy et al. 2011). The isotope ratios also indicate that the water and CO_2 were derived from the surface and atmosphere, rather than the mantle, but were in only poor communication with the atmosphere while the carbonates were precipitating. The most likely explanation appears to be precipitation in an ephemeral aquifer at a depth of a few to a few tens of meters in the regolith. Thus, the carbonate content of ALH 84001 is consistent with, although it does not require, a wet, or at least moist, Noachian climate in which shallow subsurface groundwater remained above the freezing point.

Radioisotope analyses of other Martian meteorites have given ages from as early as ~2 Gy to as recent as 150 My. However, Bouvier et al. (2005) obtained conflicting results for the shergottite meteorites Zagami and Shergotty: Sm-Nd and Lu-Hf ages were 155–185 Ma, but Pb-Pb ages were 4.05 Gy. The latter age is similar to that of ALH 84001, above. The Bouvier et al. (2005) isotope data suggest that the younger ages result from resetting of the Sm-Nd and Lu-Hf clocks by interactions with groundwater percolating through the rock, and that the Pb-Pb age represents the actual crystallization age of the rock. Thus, although these conditions may have been highly episodic, the Sm-Nd and Lu-Hf ages indicate percolation of groundwater on Mars as recently as ~150 million years ago.

The only in situ search for life so far conducted on the surface of Mars was performed by the *Viking* landers in 1976. Each of the two landers carried a set of four experiments:

1. *Gas Chromatograph–Mass Spectrometer* (*GCMS*): A soil sample was gradually heated to as high as 500° C, and the gases released were identified by measuring their molecular masses. The GCMS was not looking for biological activity, simply for organic matter regardless of origin.

2. *Gas Exchange experiment* (*GEx*): Nutrients and water were added to soil samples in a sealed chamber at ~10° C, and the air in the chamber was monitored for evolved gases such as O_2, CO_2, N_2, H_2, and CH_4 that might indicate biological activity. Control samples were preheated to sterilize them.

3. *Labeled Release* (*LR*) experiment: ^{14}C-tagged nutrients were added to soil samples at ~10° C, and radiation detectors looked for radioactive $^{14}CO_2$ that could have been released by biological activity. Control samples were preheated to sterilize them.

4. *Pyrolytic Release experiment* (*PR*): A soil sample was exposed to a Mars-like atmosphere containing $^{14}CO_2$ and ^{14}CO, and the soil was later heated to >600°C. Detection of organic volatiles containing ^{14}C would indicate biological activity. Control samples were preheated to sterilize them. Of the three biology experiments, the PR experiment was performed under the most Martian-like conditions, because it did not require either heat or liquid water during the incubation period, although these could have been provided on command.

The GCMS detected only two organic compounds in the soil in quantities above its very stringent detection limits. These were trace quantities of chloromethane (CH_3Cl) at 15 ppb at *Viking* 1 and dichloromethane (CH_2Cl_2) at 0.04–40 ppb at *Viking* 2 (Navarro-González et al. 2010). The absence of organic matter was surprising, because even lunar soils contain detectable amounts of organic compounds from meteorites. It was suggested that organic matter, at least on the Martian surface, is destroyed by oxidizing compounds present in the soil.

The CH_3Cl and CH_2Cl_2 were initially believed to be contaminants introduced by the cleaning process before launch. However, in 2008, the *Phoenix* lander detected perchlorate (ClO_4^-) in the Martian soil at a level of 0.4–0.6 wt%, thought to be in the form of magnesium perchlorate, $Mg(ClO_4)_2$, or calcium perchlorate, $Ca(ClO_4)_2$. Subsequent experiments (Navarro-González et al. 2010) were performed on Earth in which 1 wt% $Mg(ClO_4)_2$ was added to samples of Mars-like soil from the Atacama Desert in Chile containing 32 ± 6 ppm of organic carbon, and the soil was heated with a similar procedure to the *Viking* experiments, with an overlapping but extended temperature range. During the heating, the perchlorate reacted with the organics in the soil, converting them all to H_2O and CO_2 with trace amounts of CH_3Cl and CH_2Cl_2. Thus, perchlorate in the Martian soil could also have destroyed organic material when heated in the *Viking* chambers. From these experiments, Navarro-González et al. (2010) have re-interpreted the *Viking* GCMS results as a positive detection of organic carbon, with $\leq$0.1 % perchlorate and 1.5–6.5 ppm organic carbon at landing site 1 and $\leq$0.1 % perchlorate and 0.7–2.6 ppm organic carbon at landing site 2.

In the GEx experiment, quantities of O_2 were produced, but the results were considered to be consistent with non-biological chemical reactions involving compounds such as superoxides or hydrogen peroxide (Oyama and Berdahl 1977).

The PR experiment showed that very small quantities of ^{14}C were incorporated into the soil, but the agent is thought to be non-biological because it was not destroyed by heat in control tests. Subsequent experiments on Earth produced similar responses when iron-rich minerals were exposed to $^{14}CO_2$ and ^{14}CO (Klein et al. 1992).

The LR results are the most controversial. At both sites there was a strong response when nutrients were added the first time, and diminished responses on subsequent additions. These results were independent of whether the sample was from an exposed location or from a sheltered spot under a rock. Unprocessed samples were also stored in the dark for several months at ~10°C and then tested; these produced only very weak responses (<10 % of the earlier, initial responses).

The philosophy behind the delayed testing is that chemical agents are unlikely to be destroyed by temperatures of only ~10°C, whereas biological agents could be sensitive to temperatures significantly above those of their natural habitat (−21°C to −84°C).

The LR investigative team regarded the LR results as having met the criteria established for a positive result; i.e., as being consistent with the presence of living organisms in the Martian soil (Levin and Straat 1979). Others, however, have suggested that the responses could be produced by chemical reactions involving hydrogen peroxide in the soil, or other peroxides, superoxides or ozonides.

The *Viking* results are thus considered inconclusive at present.

It may be helpful, in this ambiguous situation, to consider the forms into which life on Earth (the only life we know) has evolved. Life on the Earth is based on carbon, and can be divided into *heterotrophs*, which obtain their carbon from organic molecules, and *autotrophs*, which obtain their carbon from inorganic molecules. The latter can therefore live and grow in the absence of organic matter. [*Organic molecules* are molecules that contain carbon, with certain exceptions such as simple oxides of carbon (e.g., CO and CO_2) that are considered inorganic.] Heterotrophs include all animals and most bacteria. *Phototrophs* use energy from sunlight to fix carbon, and include plants and photosynthetic bacteria. On Earth, almost all phototrophs obtain their carbon from CO_2, and are therefore autotrophs. *Chemotrophs* obtain energy from oxidation-reduction (*redox*) reactions in which donor molecules provide electrons that are passed through electron transport chains to acceptor molecules. For *chemoorganotrophs*, the donor molecules are organic, and for *chemolithotrophs* the donor molecules are inorganic, usually minerals. Many chemolithotrophs are *extremophiles*, growing in extreme environments such as hot, hydrothermal vents deep underwater. In *aerobic* metabolism, the final electron acceptor is oxygen (O_2), and the organism requires O_2 to survive; *anaerobic* metabolism does not require oxygen.

We know that the Martian surface currently is hostile to a great many life forms: the atmosphere is anoxic (lacking oxygen), the surface is highly oxidizing, there are high levels of both UV and particle radiation (because of the thin atmosphere, the lack of ozone, and the lack of a global magnetic field), and aqueous conditions have become more acidic with time through the Hesperian and Amazonian epochs (Sect. 9.3.2). Thus, if life has existed on Mars within the last 2–3 Gy, it is likely to have been anaerobic and located below the surface, and therefore chemotrophic—not impossible requirements, because bacteria have been found up to 5.3 km below the surface on Earth.

Nixon et al. (2013) and McMahon et al. (2013) provide useful insights on possible Martian metabolic processes and habitable environments, and we refer the reader to these reviews for discussion and references. Some points are summarized below.

Martian rocks are enriched in Fe and S compared to Earth, and these are known to act as donor molecules on Earth. Sulfate and ferric oxide deposits possibly associated with ancient hydrothermal or fumarolic processes are known from Gusev crater (the location of *Spirit* rover). Martian rocks are also rich in olivine and pyroxene, which can provide Fe^{2+} as an electron donor, creating Fe^{3+}. Other microbes could convert Fe^{3+} back to Fe^{2+} by using the Fe^{3+} as acceptors and hydrocarbons, H_2, or elemental sulfur as donors. Most terrestrial iron-oxidizing microbes use oxygen as the electron-acceptor, or some use nitrate (NO_3^-), but both of these are in short supply on Mars. Perchlorate (ClO_4^-), which is known to occur in Martian soils, has been suggested as an electron acceptor, but this has yet to be tested. CO_2 from the Martian atmosphere is available in abundance for carbon fixation, and basalt can provide C, H, N, O, P, and S for nutrients.

Methanogenic (methane-producing) metabolisms are also possible, and traces of methane (several parts per billion) were detected in the Martian atmosphere by orbiting spacecraft. However, *Curiosity* rover's inability to detect methane on the surface in December, 2012, argues against the existence of methanogenic microbes at the present time.

A possible subsurface habitat where conditions could be favorable for Martian microbes is vesicular basalt. Pore closure from viscous deformation is estimated to occur ~4–5 km below the Martian surface, although filling of pores by precipitates could reduce this depth. Estimates for the average global thickness of the Martian cryosphere ("cold sphere," where water would be permanently frozen) range from ~3 to ~9 km below the surface, and the water table could be anywhere from directly below the cryosphere to several kilometers depth, depending on the amount of water available. Subsurface ice is known to exist today to latitudes as close as 25° from the equator (Sect. 9.3.3). If this ice extends to the base of the cryosphere, then geothermal heat could maintain liquid water.

It has also been suggested that Martian lava tubes could have supported life in the past, perhaps even continuing to the present. Popa et al. (2012) investigated a lava tube in the Cascade Mountains of Oregon, U.S.A., as a possible analogue of Martian lava tubes, isolating 29 microbial strains in samples taken from a dark, moist, perennial ice habitat near 0 °C at the basalt–ice interface. On the basis of initial experiments, one of these, a strain of *Pseudomonas*, was selected for more thorough study. It is aerobic, but in a mineral medium with olivine it was able to oxidize Fe^{2+} from olivine as the only electron donor and O_2 as the electron acceptor. It can also obtain carbon from CO_2.

Thus, while we have relatively little information on the Martian subsurface environment, the existence of subsurface microbial life on Mars now or in the past is not ruled out. Indeed, according to Lunine (2005, p. 317), some sources suggest that Earth's *endolithic* biosphere, living in subsurface rock, may exceed the mass of all other life forms on Earth. Even if this controversial estimate is excessive, the subsurface biomass is likely to be substantial.

It might also be worth noting that, although some microbes can process perchlorates, the levels found in the Martian soil by the Phoenix lander (~0.5 wt%) would be toxic to humans. Future explorers on Mars, therefore, would need to take special precautions to shield themselves from perchlorate contamination.

This concludes our examination of the surfaces and interiors of the terrestrial planets. The atmospheres of planets will be considered in Chap. 10, and their ionospheres and magnetospheres in Chap. 11, of Milone and Wilson 2014.

Challenges

[9.1] The root-mean-square (rms) speed, v_{rms}, of atoms or molecules of mass m in a gas in thermal equilibrium at temperature T is given by the equation 1/2 m $(v_{\mathrm{rms}})^2$ = (3/2)kT, where k is Boltzmann's constant (Milone & Wilson 2014, Sect. 10.1). If these atoms or molecules are a component of a planetary atmosphere, then this component escapes into space over a timescale of weeks if v_{rms} = 1/3 v_{esc}, where v_{esc} is the escape velocity from the planet [equation (5.39)]; 10^4 years if v_{rms} = 1/4 v_{esc}, and 10^8 years if v_{rms} = 1/5 v_{esc}, Calculate (a) the escape velocity of an atom from the surface of Mercury and (b) the rms speeds of atoms of S and Fe vapor in thermal equilibrium with the surface of a magma ocean at a temperature of (i) 1,000 K; (ii) 1,500 K on the planet Mercury. (c) For each of these temperatures, comment on the retention of these atoms by Mercury if the ocean surface is molten for 10^5 years.

[9.2] Compute the impact speed and specific impact energy for an asteroid colliding with each of the terrestrial planets and the Earth's moon. Assume the asteroid to have an orbit with the same semi-major axis as the orbit of the planet.

[9.3] Ignore atmospheric effects for the situation in [9.2] and comment on the size of the craters one would expect for each body for the same mass of impactor of a stony meteorite (say $\rho = 3{,}500$ kg/m^3). Is it reasonable to suppose one should ignore atmospheric effects?

[9.4] Estimate the mass of impactor required to create the Hellas basin on Mars. Show all reasoning.

[9.5] Compare the observed and global equilibrium temperature of Venus. Is it reasonable to ignore internal heat sources on this planet? If the only source of heat on Venus were internal, compute its equilibrium temperature.

[9.6] Solar evolution models suggest that the Sun will be 10 % more luminous ~1 Gy from now. If it were so, how would this affect the environments of Earth and Mars at that stage?

References

Abell, G.: Exploration of the Universe, 2nd edn. Holt, Rinehart and Winston, New York (1969)

Andrews-Hanna, J., Zuber, M.T., Banerdt, W.B.: The Borealis basin and the origin of the Martian crustal dichotomy. Nature **453**, 1212–1215 (2008)

Blewett, D.T., Chabot, N.L., Denevi, B.W., Ernst, C.M., Head, J.W., Izenberg, N.R., Murchie, S.L., Solomon, S.C., Nittler, L.R., McCoy, T.J., Xiao, Z., Baker, D.M.H., Fassett, C.I., Braden, S.E., Oberst, J., Scholten, F., Preusker, F., Hurwitz, D.M.: Hollows on Mercury: MESSENGER evidence for geologically recent volatile-related activity. Science **333**, 1856–1859 (2011)

Bondarenko, N.V., Head, J.W., Ivanov, M.A.: Present-day volcanism on Venus: evidence from microwave radiometry. Geophys. Res. Lett. **37**, L23202 (2010) (5 pages)

Borg, L.E., Connelly, J.N., Nyquist, L.E., Shih, C.-Y., Wiesmann, H., Reese, Y.: The age of the carbonates in Martian meteorite ALH84001. Science **286**, 90–94 (1999)

Bouvier, A., Blichert-Toft, J., Vervoort, J.D., Albarède, F.: The age of SNC meteorites and the antiquity of the Martian surface. Earth Planet. Sci. Lett. **240**, 221–233 (2005)

Bradley, J.P., Harvey, R.P., McSweem Jr., H.Y.: No 'nanofossils' in Martian meteorite. Nature **390**, 454 (1997)

Bullock, M.A., Grinspoon, D.H.: The recent evolution of climate on Venus. Icarus **150**, 19–37 (2001)

Byrne, S., Dundas, C.M., Kennedy, M.R., Mellon, M.T., McEwen, A.S., Cull, S.C., Daubar, I.J., Shean, D.E., Seelos, K.D., Murchie, S.L., Cantor, B.A., Arvidson, R.E., Edgett, K.S., Reufer, A., Thomas, N., Harrison, T.N., Posiolova, L.V., Seelos, F.P.: Distribution of mid-latitude ground ice on Mars from new impact craters. Science **325**, 1674–1676 (2009)

Carr, M.H.: Mars: surface and interior. In: Weissman, P.R., McFadden, L.-A., Johnson, T.V. (eds.) Encyclopedia of the Solar System, pp. 291–308. Academic, San Diego, CA (1999)

Chabot, N.L., Ernst, C.M., Harmon, J.K., Murchie, S.L., Solomon, S.C., Blewett, D.T., Denevi, B.W.: Craters hosting radar-bright deposits in mercury's north polar region, 43rd Lunar and planetary science conference, Abstract 1476 (2012)

Chapman, M.G., Neukum, G., Dumke, A., Michael, G., van Gasselt, S., Kneissl, T., Zuschneid, W., Hauber, E., Ansan, V., Mangold, N., Masson, P.: Noachian–Hesperian geologic history of the Echus Chasma and Kasei Valles system on Mars: new data and interpretations. Earth Planet. Sci. Lett. **294**, 256–271 (2010)

Charlier, B., Grove, T.L., Zuber, M.T.: Composition and differentiation of 'basalts' at the surface of mercury. 43rd Lunar and planetary science conference, Abstract 1400 (2012)

Chevrier, V.F., Rivera-Valentin, E.G.: Formation of recurring slope lineae by liquid brines on present-day Mars. Geophys. Res. Lett. **39** (2012). doi: 10.1029/2012GL054119

Chevrier, V., Dehouck, E., Gaudin, A., Mangold, N., Mathe, P.E., Rochette, P.: Experimental verification of the "burns" hypothesis for the formation of meridiani planum sediments through weathering of sulfide-rich deposits. 41st Lunar and planetary science conference, Abstract 2440 (2010)

Consolmagno, G.J., Schaefer, M.W.: Worlds Apart: A Textbook in the Planetary Sciences. Prentice Hall, Englewood Cliffs, NJ (1994)

Cox, A.N. (ed.): Allen's Astrophysical Quantities, 4th edn. Springer, New York (2000)

Dauphas, N., Pourmand, A.: Hf–W–Th evidence for rapid growth of Mars and its status as a planetary embryo. Nature **473**, 489–493 (2011)

Denevi, B.W., Robinson, M.S., Solomon, S.C., Murchie, S.L., Blewett, D.T., Domingue, D.L., McCoy, T.J., Ernst, C.M., Head, J.W., Watters, T.R., Chabot, N.L.: The evolution of mercury's crust: a global perspective from MESSENGER. Science **324**, 613–618 (2009)

Di Achille, G., Popa, C., Massironi, M., Ferrari, S., Giacomini, L., Mazzotta Epifani, E., Pozzobon, R., Zusi, M., Cremonese, G., and Palumbo, P.: Mapping Mercury's tectonic features at the terminator: implications for radius change estimates and thermal history models. 43rd Lunar and planetary science conference, Abstract 2176 (2012)

Dicke, R.H., Goldenberg, H.M.: Solar oblateness and general relativity. Phys. Rev. Lett. **18**, 313–316 (1967)

Dickson, L., Head, J.W., Whitten, J.L., Fassett, C.I., Neumann, G.A., Smith, D.E., Zuber, M. T., Phillips, R.J.: Topographic rise in the northern smooth plains of Mercury: characteristics from MESSENGER image and altimetry data and candidate modes of origin. 43rd Lunar and planetary science conference, Abstract 2249 (2012)

Diniega, S., Hansen, C.J., McElwaine, J.N., Hugenholtz, C.H., Dundas, C.M., McEwen, A.S., Bourke, M.C.: A new dry hypothesis for the formation of Martian linear gullies. Icarus **225**, 526–537 (2013)

Dombard, A.J., Hauck, S.A., Solomon, S.C., Phillips, R.J.: Potential for long-wavelength folding on Mercury. 32nd Lunar and planetary science conference. Abstract 2035 (2001)

Elkins-Tanton, L.T., Hess, P.C., Parmentier, E.M.: Possible formation of ancient crust on Mars through magma ocean processes. J. Geophys. Res. **93**, E12S01 (2005). doi: 10.1029/2005JE002480

Erkeling, G., Reiss, D., Hiesinger, H., Jaumann, R.: Morphologic, stratigraphic and morphometric investigations of valley networks in Eastern Libya Montes, Mars: implications for the Noachian/Hesperian climate change. Earth Planet. Sci. Lett. **294**, 291–305 (2010)

Esposito, L.W., Copley, M., Eckert, R., Gates, L., Stewart, A.I.F., Worden, H.: Sulfur dioxide at the Venus cloud tops, 1978–1986. J. Geophys. Res. **93**, 5267–5276 (1988)

Fegley Jr., B., Prinn, R.G.: Estimation of the rate of volcanism on Venus from reaction rate measurements. Nature **337**, 55–58 (1989)

Fegley Jr., B., Klinglehöfer, G., Lodders, K., Widemann, T.: Geochemistry of surface-atmosphere interactions on Venus. In: Bougher, S.W., Hunten, D.M., Phillips, R.J. (eds.) Venus II: Geology, Geophysics, Atmosphere, and Solar Wind Environment, pp. 591–636. University of Arizona Press, Tucson, AZ (1997)

Fivian, M.D., Hudson, H.S., Lin, R.P., Zahid, H.J.: A large excess in apparent solar oblateness due to surface magnetism. Science **322**, 560–562 (2008)

Folkner, W.M., Yoder, C.F., Yuan, D.N., Standish, E.M., Preston, R.A.: Interior structure and seasonal mass redistribution of Mars from radio tracking of Mars pathfinder. Science **178**, 1749–1751 (1997)

Frey, H.V.: Impact constraints on, and a chronology for, major events in early Mars history. J. Geophys. Res. **111**, E08S91 (2006). doi: 10.1029/2005JE002449

Fuller, E.R., Head III, J.W.: Amazonis Planitia: the role of geologically recent volcanism and sedimentation in the formation of the smoothest plains on Mars. J. Geophys. Res. **107**, 5081 (2002). doi: 10.1029/2002JE001842 (25 pages)

Gardner, P.M., Robins, B.: The Olivine-Plagioclase reaction: geological evidence from the Sieland Petrographic Province, Northern Norway. Contrib. Mineral. Petrol. **44**, 149–156 (1974)

Glaze, L.S., Stofan, E.R., Smrekar, S.E., Baloga, S.M.: Insights into Corona formation through statistical analyses. J. Geophys. Res. **107**, 5135–5146 (2002)

Golden, D.C., Ming, D.W., Schwandt, C.S., Lauer, H.V., Socki, R.A., Morris, R.V., Lofgren, G.E., McKay, G.A.: A simple inorganic process for formation of carbonates, magnetites, and sulfides in Martian meteorite ALH84001. Am. Mineral. **86**, 370–375 (2001)

Halevy, I., Fischer, W.W., Eiler, J.M.: Carbonates in the Martian meteorite Allan Hills 84001 formed at 18 ± 4 °C in a near-surface aqueous environment. Proc. Natl. Acad. Sci. **108**, 16895–16899 (2011)

Hansen, V.L.: LIPs on Venus. Chem. Geol. **241**, 354–374 (2007)

Hansen, V.L., López, I.: Venus records a rich early history. Geology **38**, 311–314 (2011)

Harmon, J.K., Slade, M.A., Rice, M.S.: Radar imagery of Mercury's putative polar ice: 1999–2005 Arecibo results. Icarus **211**, 37–50 (2011)

Head, J.W., Basilevsky, A.T.: Venus: surface and interior. In: Weissman, P.R., McFadden, L.-A., Johnson, T.V. (eds.) Encyclopedia of the Solar System, pp. 161–189. Academic, San Diego, CA (1999)

Head, J.W., Chapman, C.R., Strom, R.G., Fassett, C.I., Denevi, B.W., Blewett, D.T., Ernst, C.M., Watters, T.R., Solomon, S.C., Murchie, S.L., Prockter, L.M., Chabot, N.L., Gillis-Davis, J.J., Whitten, J.L., Goudge, T.A., Baker, D.M.H., Hurwitz, D.M., Ostrach, L.R., Xiao, Z., Merline, W.J., Kerber, L., Dickson, J.L., Oberst, J., Byrne, P.K., Klimczak, C., Nittler, L.R.: Flood volcanism in the Northern high latitudes of Mercury revealed by MESSENGER. Science **333**, 1853–1855 (2011)

Hoogenboom, T., Houseman, G.A.: Rayleigh–Taylor instability as a mechanism for corona formation on Venus. Icarus **180**, 292–307 (2006)

Ivanov, M.A., Head, J.W.: Global geological map of Venus. Planet. Space Sci. **59**, 1559–1600 (2011)

Jull, A.J.T., Courtney, C., Jeffrey, D.A., Beck, J.W.: Isotopic evidence for a terrestrial source of organic compounds found in Martian meteorites Allan Hills84001 and Elephant Moraine 79001. Science **279**, 366–369 (1997)

Kelley, D.H., Milone, E.F.: Exploring Ancient Skies, 2nd edn. Springer, New York (2011)

Kieffer, H.H., Christenson, P.R., Titus, T.N.: CO_2 jets formed by sublimation beneath translucent slab ice in Mars' seasonal south polar ice cap. Nature **442**, 793–796 (2006)

Klein, H.P., Horowitz, N.H., Biemann, K.: The search for extant life on Mars. In: Keiffer, H.H., Jakosky, B.M., Snyder, C.W., Matthews, M.S. (eds.) Mars, pp. 1221–1233. University of Arizona Press, Tucson, AZ (1992)

Konopliv, A.S., Asmar, S.W., Folkner, W.M., Karatekin, Ö., Nunes, D.C., Smrekar, S.E., Yoder, C.F., Zuber, M.T.: Mars high resolution gravity fields from MRO, Mars seasonal gravity, and other dynamical parameters. Icarus **211**, 401–428 (2011)

Krassilnikov A.S., Head, J.W.: Arachnoids on Venus: structural analysis, classification and models of formation. 34th Lunar and planetary science conference, Abstract 1220 (2003)

Kreslavsky, M.A., Head, J.W.: Fate of outflow effluents in the northern lowlands of Mars: the Vastitas Borealis formation as a sublimation residue from frozen ponded bodies of water. J. Geophys. Res. **107** (2002). doi: 10.1029/2001JE001831

Lane, M.D., Christensen, P.R., Hartmann, W.K.: Utilization of the THEMIS visible and infrared imaging data for crater population studies of the Meridiani Planum landing site. Geophys. Res. Lett. **30**, 8071–8074 (2003)

Langevin, Y., Douté, S., Vincendon, M., Poulet, F., Bibring, J.-P., Gondet, B., Schnitt, B., Forget, F.: No signature of clear CO_2 ice from the 'cryptic' regions in Mars' south polar cap. Nature **442**, 790–792 (2006)

Lapen, T.J., Righter, M., Brandon, A.D., Debaille, V., Beard, B.L., Shafer, J.T., Peslier, A.H.: A younger age for ALH84001 and its geochemical link to shergottite sources in Mars. Science **328**, 347–351 (2010)

Levin, G.V., Straat, P.A.: Completion of the Viking labeled release experiment on Mars. J. Mol. Evol. **14**, 167–183 (1979)

Margot, J.L., Peale, S.J., Jurgens, R.F., Slade, M.A., Holin, I.V.: Large longitude libration of Mercury reveals a molten core. Science **316**, 710–714 (2007)

Margot, J.L., Peale, S.J., Solomon, S.C., Hauck II, S.A., Ghigo, F.D., Jurgens, R.F., Yseboodt, M., Giorgini, J.D., Padovan, S., Campbell, D.B.: Mercury's moment of inertia from spin and gravity data. J. Geophys. Res. **117**, E00L09 (2012). doi: 10.1029/2012JE004161 (11 pages)

McKay, D.S., Clemett, S.J., Gibson, E.K., Jr., Thomas-Keprta, K., Wentworth, S.J.: Are carbonate globules, magnetites, and PAHs in ALH84001 really terrestrial contaminants? Lunar Planet. Soc., **XXXIII**, Pdf. 1943 (2002)

McKay, D.S., Gibson Jr., E.K., Thomas-Keprta, K.L., Vali, H., Romanek, C.S., Clemett, S.J., Chillier, X.D.F., Maechling, C.R., Zare, R.N.: Search for past life on Mars: possible relic biogenic activity in Martian meteorite ALH84001. Science **273**, 924–930 (1996)

McKay, D.S., Gibson Jr., E., Thomas-Keprta, K.: Reply to: no 'nanofossils' in Martain meteorites. Nature **390**, 465–466 (1997)

McKinnon, W.B., Zahnle, K.J., Ivanov, B.A., Melosh, H.J.: Cratering on Venus: models and observations. In: Bougher, S.W., Hunten, D.M., Phillips, R.J. (eds.) Venus II: Geology, Geophysics, Atmosphere, and Solar Wind Environment, pp. 969–1014. University of Arizona Press, Tucson, AZ (1997)

McMahon, S., Parnell, J., Ponicka, J., Boyce, A.: The habitability of vesicles in Martian basalt. Astron. Geophys. **54**, 1.17–1.21 (2013)

McSween, H.Y., Taylor, G.J., Wyatt, M.B.: Elemental composition of the Martian crust. Science **324**, 736–739 (2009)

Mellon, M.T., Feldman, W.C., Prettyman, T.H.: The presence and stability of ground ice in the southern hemisphere of Mars. Icarus **169**, 324–340 (2004)

Meunier, A., Petit, S., Ehlmann, B.L., Dudoignon, P., Westall, F., Mas, A., El Albani, A., Ferrage, E.: Magmatic precipitation as a possible origin of Noachian clays on Mars. Nat. Geosci. **5**, 739–743 (2012)

Milone, E.F., Wilson, W.J.F.: Solar System Astrophysics: Planetary Atmospheres and the Outer Solar System, 2nd edn. Springer, New York (2014)

Misner, C.W., Thorne, K.S., Wheeler, J.A.: Gravitation. Freeman, San Francisco, CA (1973)

Moroz, V.I., Ekonomov, A.P., Moshkin, B.E., Revercomb, H.E., Sromovsky, L.A., Schofield, J.T., Spänkuch, D., Taylor, F.W., Tomasko, M.G.: Solar and thermal radiation in the Venus atmosphere. Adv. Space Res. **5**(11), 197–232 (1985)

Navarro-González, R., Vargas, E., de la Rosa, J., Raga, A.C., McKay, C.P.: Reanalysis of the Viking results suggests perchlorate and organics at midlatitudes on Mars. J. Geophys. Res. **115** (2010). doi: 10.1029/2010JE003599 (11 pages)

Neumann, G.A., Cavanagh, J.F., Sun, X., Mazarico, E., Smith, D.E., Zuber, M.T., Solomon, S. C., Paige, D.A.: Dark material at the surface of polar crater deposits on Mercury. 43rd Lunar and planetary science conference, Abstract 2651 (2012)

Nittler, L.R., Starr, R.D., Weider, S.Z., McCoy, T.J., Boynton, W.V., Ebel, D.S., Ernst, C.M., Evans, L.G., Goldsten, J.O., Hamara, D.K., Lawrence, D.J., McNutt Jr., R.L., Schlemm II, C.E., Solomon, S.C., Sprague, A.L.: The major-element composition of Mercury's surface from MESSENGER X-ray spectrometry. Science **333**, 1847–1850 (2011)

Nixon, S.L., Cousins, C.R., Cockell, C.S.: Plausible microbial metabolisms on Mars. Astron. Geophys. **54**, 1.13–1.16 (2013)

Oberst, J., Preusker, F., Phillips, R.J., Watters, T.R., Head, J.W., Zuber, M.T., Solomon, S.C.: The morphology of Mercury's Caloris basin as seen in MESSENGER stereo topographic models. Icarus **209**, 230–238 (2010)

Ojha, L., McEwen, A., Dundas, C., Mattson, S., Byrne, S., Schaefer, E., Masse, M.: Recurring slope linneae on Mars: updated global survey results. 43rd Lunar and planetary science conference, Abstract 2591 (2012)

Oyama, V.I., Berdahl, B.J.: The Viking gas exchange experiment results from Chryse and Utopia surface samples. J. Geophys. Res. **82**, 4669–4676 (1977)

Paige, D.A., Siegler, M.A., Harmon, J.K., Smith, D.E., Zuber, M.T., Neumann, G.A., Solomon, S.C.: Thermal stability of frozen volatiles in the north polar region of mercury. 43rd Lunar and planetary science conference, Abstract 2875 (2012)

Pasachoff, J.M., Sheehan, W.: A major discovery in doubt. Sky Telesc. **125**(1), 86 (2013)

Pasachoff, J.M., Schneider, G., Golub, L.: Space studies of the black-drop effect at a Mercury transit. Bull. Am. Astron. Soc. **35**, 1202 (2003)

Peplowski, P.N., Evans, L.G., Hauck II, S.A., McCoy, T.J., Boynton, W.V., Gillis-Davis, J.J., Ebel, D.S., Goldsten, J.O., Hamara, D.K., Lawrence, D.J., McNutt Jr., R.L., Nittler, L.R., Solomon, S.C., Rhodes, E.A., Sprague, A.L., Starr, R.D., Stockstill-Cahill, K.R.: Radioactive elements on Mercury's surface from MESSENGER: implications for the planet's formation and evolution. Science **333**, 1850–1852 (2011)

Phillips, R.J., Raubertas, R.F., Arvidson, R.E., Sarkar, I.C., Herrick, R.R., Izenberg, N., Grimm, R. E.: Impact craters and Venus resurfacing history. J. Geophys. Res. **97**, 15,923–15,948 (1992)

Popa, R., Smith, A.R., Popa, R., Boone, J., Fisk, M.: Olivine-respiring bacteria isolated from the rock-ice interface in a lava-tube cave, a Mars analog environment. Astrobiology **12**(1), 9–18 (2012)

Read, P.L., Lewis, S.R.: The Martian Climate Revisited. Springer, New York etc.; Praxis, Chichester (2004)

Reese, C.C., Orth, C.P., Solomatov, V.S.: Impact megadomes and the origin of the martian crustal dichotomy. Icarus **213**, 433–442 (2011)

Rivoldini, A., Van Hoolst, T., Verhoeven, O., Mocquet, A., Dehant, V.: Geodesy constraints on the interior structure and composition of Mars. Icarus **213**, 451–472 (2011)

Robinson, M.S., Taylor, G.J.: Ferrous oxide in Mercury's crust and mantle. Meteorit. Planet. Sci. **36**, 841–847 (2001)

Ross, F.E.: Photographs of Venus. Astrophys. J. **68**, 57–92 (1928)

Segura, T.L., Toon, O.B., Colaprete, A., Zahnle, K.: Environmental effects of large impacts on Mars. Science **298**, 1977–1980 (2002)

Segura, T.L., Toon, O.B., Colaprete, A.: Modeling the environmental effects of moderate-sized impacts on Mars. J. Geophys. Res. **113**, E11007 (2008)

Shiltsev, V.: Mikhail Lomonosov and the dawn of Russian science. Phys. Today **65**(2), 40–46 (2012)

Smith, D.E., Zuber, M.T., Phillips, R.J., Solomon, S.C., Hauck II, S.C., Lemoine, F.G., Mazarico, E., Neumann, G.A., Peale, S.J., Margot, J.-L., Johnson, C.L., Torrence, M.H., Perry, M.E., Rowlands, D.D., Goossens, S., Head, J.W., Taylor, A.H.: Gravity field and internal structure of Mercury from MESSENGER. Science **336**, 214–217 (2012)

Smrekar, S.E., Stofan, E.R.: Corona formation and heat loss on Venus by coupled upwelling and delamination. Science **277**, 1289–1294 (1997)

Smrekar, S.E., Stofan, E.R., Mueller, N., Treiman, A., Elkins-Tanton, L., Helbert, J., Piccioni, G., Drossart, P.: Recent hotspot volcanism on Venus from VIRTIS emissivity data. Science **328**, 605–608 (2010)

Squyres, S.W., Janes, D.M., Baer, G., Bindschadler, D.L., Schubert, G., Sharpton, V.L., Stofan, E.R.: The morphology and evolution of coronae on Venus. J. Geophys. Res. **97**, 13,611–13,634 (1992)

Squyres, S.W., Arvidson, R.E., Bollen, D., Bell III, J.F., Brückner, J., Cabrol, N.A., Calvin, W.M., Carr, M.H., Christensen, P.R., Clark, B.C., Crumpler, L., Des Marais, D.J., d'Uston, C., Economou, T., Farmer, J., Farrand, W.H., Folkner, W., Gellert, R., Glotch,T.D., Golombek, M., Gorevan, S., Grant, J.A., Greeley, R., Grotzinger, J., Herkenhoff, K.E., Hviid, S., Johnson, J.R., Klingelhöfer, G., Knoll, A.H., Landis, G., Lemmon, M., Li, R., Madsen, M.B., Malin, M. C., McLennan, S.M., McSween, H.Y., Ming, D.W., Moersch, J., Morris, R.V., Parker, T., Rice Jr., J.W., Richter, L., Rieder, R., Schröder, C., Sims, M., Smith, M., Smith, P., Soderblom, L.A., Sullivan, R., Tosca, N.J., Wänke, H., Wdowiak, T., Wolff, M., Yen, A.: Overview of the Opportunity Mars Exploration Rover Mission to Meridiani Planum: Eagle Crater to Purgatory Ripple. J. Geophys. Res **111**, (2006). doi:10.1029/2006JE002771

Stevenson, D.J., Spohn, T., Schubert, G.: Magnetism and thermal evolution of the terrestrial planets. Icarus **54**, 466–489 (1983)

Stofan, E.R., Hamilton, V.E., Janes, D.M., Smrekar, S.E.: Coronae on Venus: morphology and origin. In: Bougher, S.W., Hunten, D.M., Phillips, R.J. (eds.) Venus II: Geology, Geophysics, Atmosphere, and Solar Wind Environment, pp. 931–965. University of Arizona Press, Tucson, AZ (1997)

Strom, R.G., Sprague, A.L.: Exploring Mercury: The Iron Planet. Springer-Verlag, Berlin; Praxis, Chichester, UK (2003)

Surkov, Y.A.: Studies of Venus rocks by Veneras 8, 9, and 10. In: Hunten, D.M., Colin, L., Donahue, T.M., Moroz, V.I. (eds.) Venus, pp. 154–158. University of Arizona Press, Tucson, AZ (1983)

Taylor, S.R.: Solar System Evolution: A New Perspective. University Press, Cambridge (1992)

Taylor, G.J., Scott, E.R.D.: Mercury: an end-member planet or a cosmic accident? Mercury: space environment, surface, and interior, Abstract 8065 (2001)

Thomas-Keprta, K.L., Clemett, S.J., Bazylinski, D.A., Kirschvink, J.L., McKay, D.S., Wentworth, S.J., Vali, H., Gibson Jr., E.K., Romanek, C.S.: Magnetofossils from ancient Mars: a robust biosignature in the Martian meteorite ALH84001. Appl. Environ. Microbiol. **68**(8), 3663–3672 (2002)

Toon, O.B., Segura, T., Zahnle, K.: The formation of Martian river valleys by impacts. Annu. Rev. Earth Planet. Sci. **38**, 303–322 (2010)

Treiman, A.H.: Submicron magnetite grains and carbon compounds in Martian meteorite ALH84001: inorganic, abiotic formation by shock and thermal metamorphism. Astrobiology **3**, 369–392 (2003)

Treiman, A.H., Amundsen, H.E.F., Blake, D.F., Bunch, T.: Hydrothermal origin for carbonate globules in Martian meteorite ALH84001: a terrestrial analogue from Spitzbergen (Norway). Earth Planet. Sci. Lett. **204**, 323–332 (2002)

Turcotte, D.L., Morein, G., Roberts, D., Malamud, B.D.: Catastrophic resurfacing and episodic subduction on Venus. Icarus **139**, 49–54 (1999)

Vaughan, W.M., Helbert, J., Blewett, D.T., Head, J.W., Murchie, S.L., Gwinner, K., McCoy, T.J., Solomon, S.C.: Hollow-forming layers in impact craters on Mercury: massive sulfide or chloride deposits formed by impact melt differentiation?" 43rd Lunar and planetary science conference, Abstract 1187 (2012)

Veasey, M., Dumberry, M.: The influence of Mercury's inner core on its physical libration. Icarus **214**, 265–274 (2011)

Vincendon, M., Mustard, J., Forget, F., Kreslavsky, M., Spiga, A., Murchie, S., Bibring, J.-P.: Near-tropical subsurface ice on Mars. Geophys. Res. Lett. **37** (2010). doi: 10.1029/2009GL041426 (5 pages)

Watters, T.R., McGovern, P.J., Irwin, R.P.: Hemispheres apart: the crustal dichotomy on Mars. Annu. Rev. Earth Planet. Sci. **35**, 621–652 (2007)

Watters, T.R., Solomon, S.C., Robinson, M.S., Head, J.W., André, S.L., Hauck, S.A., Murchie, S. L.: The tectonics of Mercury: the view after MESSENGER's first flyby. Earth Planet. Sci. Lett. **285**, 283–296 (2009)

Williams, R.M.E, Dietrich, W.E., Grotzinger, J.P., Gupta, S., Malin, M.C., Palucis, M.C., Rubin, D., Stack, K., Sumner, D.Y., Yingst, Y., Bridges, J.C., Goetz, W., Koefoed, A., Jensen, J.K., Madsen, M.B., Schwenzer, S.P., Deen, R.G., Pariser, O., The MSL Science Team.: Curiosity's MastCam images reveal conglomerate outcrops with water-transported pebbles. 44th Lunar and planetary science conference, Abstract 1617 (2013)

Zuber, M.T.: The crust and mantle of Mars. Nature **412**, 220–227 (2001)

Zuber, M.T., Smith, D.E., Phillips, R.J., Solomon, S.C., Neumann, G.A., Hauck II, S.A., Peale, S.J., Barnouin, O.S., Head, J.W., Johnson, C.L., Lemoine, F.G., Mazarico, E., Sun, X., Torrence, M.H., Freed, A.M., Klimczak, C., Margot, J.-L., Oberst, J., Perry, M.E., McNutt Jr., R.L., Balcerski, J.A., Michel, N., Talpe, M.J., Yang, D.: Topography of the Northern hemisphere of Mercury from MESSENGER laser altimetry. Science **336**, 217–220 (2012)

Index

E.F. Milone and W.J.F. Wilson, *Solar System Astrophysics: Background Science and the Inner Solar System*, Astronomy and Astrophysics Library,
DOI 10.1007/978-1-4614-8848-4,