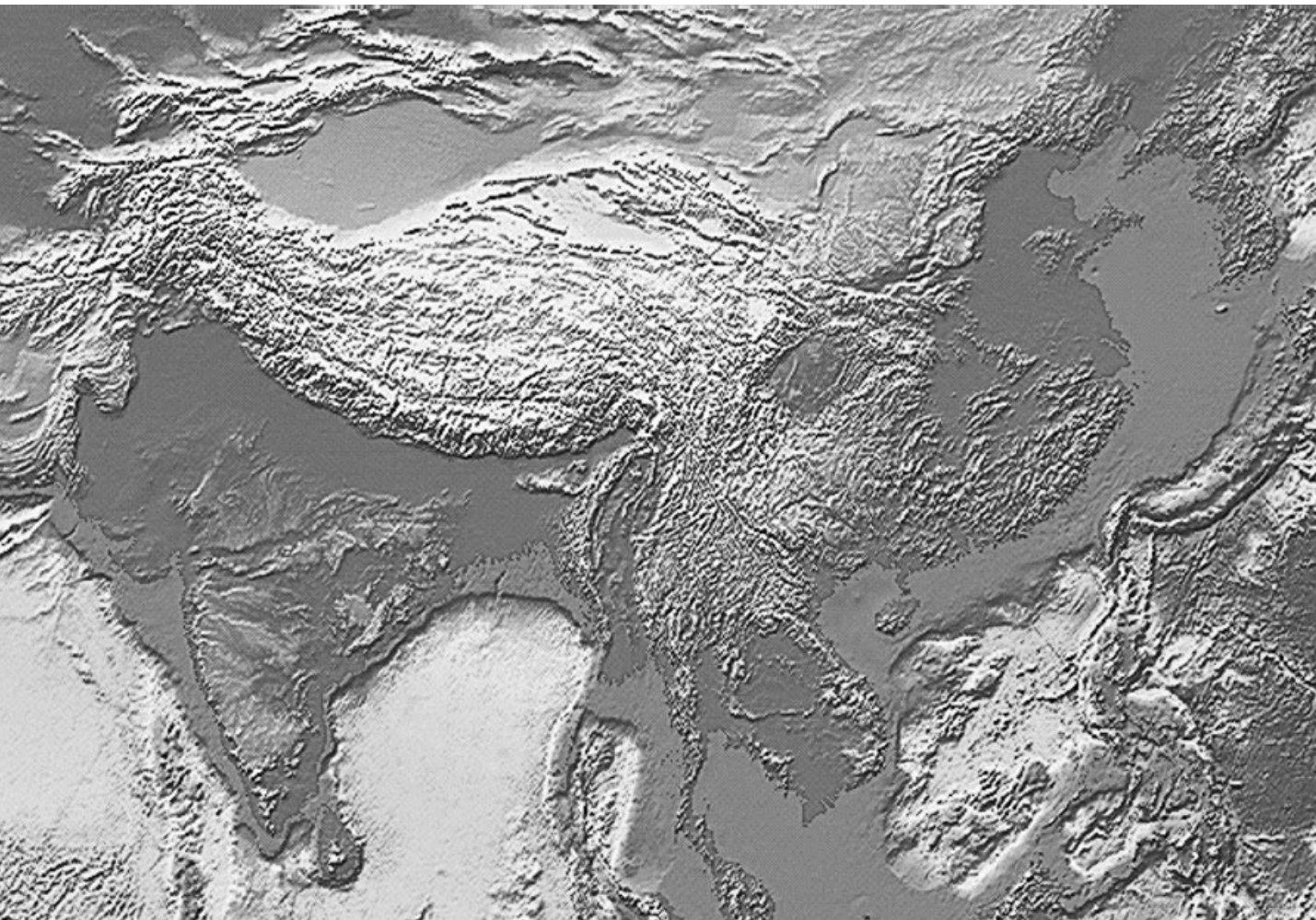
2 The interior of the Earth



2.1 EARTHQUAKE SEISMOLOGY

2.1.1 Introduction

Much of our knowledge of the internal constitution of the Earth has come from the study of the seismic waves generated by earthquakes. These waves follow various paths through the interior of the Earth, and by measuring their travel times to different locations around the globe it is possible to determine its large-scale layering. It is also possible to make inferences about the physical properties of these layers from a consideration of the velocities with which they transmit the seismic waves.

2.1.2 Earthquake descriptors

Earthquakes are normally assumed to originate from a single point known as the *focus* or *hypocenter* (Fig. 2.1), which is invariably within about 700 km of the surface. In reality, however, most earthquakes are generated by movement along a fault plane, so the focal region may extend for several kilometers. The point on the Earth's surface vertically above the focus is the *epicenter*. The angle subtended at the center of the Earth by the epicenter and the point at which the seismic waves are detected is known as the *epicentral angle* Δ. The *magnitude* of an earthquake is a measure of its energy release on a logarithmic scale; a change in magnitude of one on the *Richter* scale implies a 30-fold increase in energy release (Stein & Wysession, 2003).

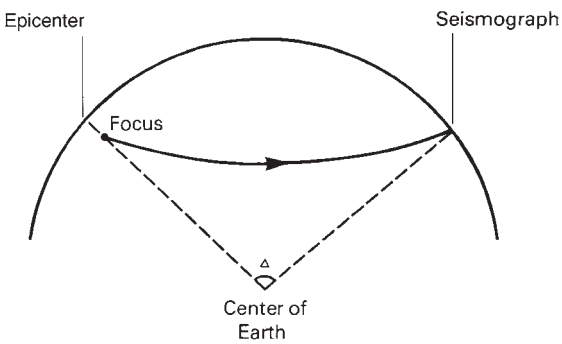


Figure 2.1 *Illustration of epicentral angle* Δ.

2.1.3 Seismic waves

The strain energy released by an earthquake is transmitted through the Earth by several types of seismic wave (Fig. 2.2), which propagate by elastic deformation of the rock through which they travel. Waves penetrating the interior of the Earth are known as *body waves*, and consist of two types corresponding to the two possible ways of deforming a solid medium. *P waves*, also known as *longitudinal* or *compressional* waves, correspond to elastic deformation by compression/dilation. They cause the particles of the transmitting rock to oscillate in the direction of travel of the wave so that the disturbance proceeds as a series of compressions and rarefactions. The velocity of a P wave V_p is given by:

$$V_p = \sqrt{\frac{k + \frac{4}{3}\mu}{\rho}}$$

where k is the bulk modulus, μ the shear modulus (rigidity), and ρ the density of the transmitting medium. *S waves*, also known as *shear* or *transverse* waves, correspond to elastic deformation of the transmitting medium by shearing and cause the particles of the rock

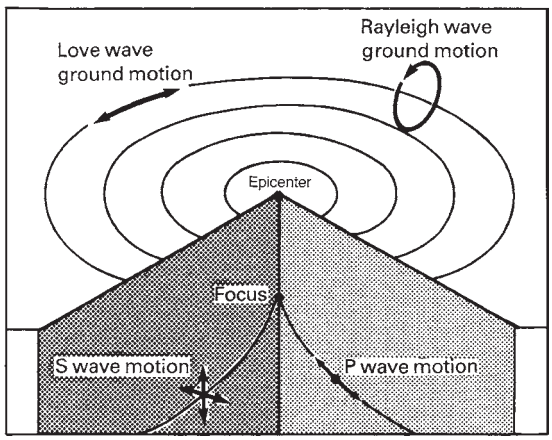


Figure 2.2 *Focus and epicenter of an earthquake and the seismic waves originating from it (after Davies, 1968, with permission from Iliffe Industrial Publications Ltd).*

to oscillate at right angles to the direction of propagation. The velocity of an S wave V_s is given by:

$$V_s = \sqrt{\frac{\mu}{\rho}}$$

Because the rigidity of a fluid is zero, S waves cannot be transmitted by such a medium.

A consequence of the velocity equations for P and S waves is that the P velocity is about 1.7 times greater than the S velocity in the same medium. Consequently, for an identical travel path, P waves arrive before S waves. This was recognized early in the history of seismology, and is reflected in the names of the body waves (P is derived from primus and S from secundus). The passage of body waves through the Earth conforms to the laws of geometric optics in that they can be both refracted and reflected at velocity discontinuities.

Seismic waves whose travel paths are restricted to the vicinity of a free surface, such as the Earth's surface, are known as *surface waves*. *Rayleigh waves* cause the particles of the transmitting medium to describe an ellipse in a vertical plane containing the direction of propagation. They can be transmitted in the surface of a uniform half space or a medium in which velocity changes with depth. *Love waves* are transmitted whenever the S wave velocity of the surface layer is lower than that of the underlying layer. Love waves are essentially horizontally polarized shear waves, and propagate by multiple reflection within this low velocity layer, which acts as a wave guide.

Surface waves travel at lower velocities than body waves in the same medium. Unlike body waves, surface waves are dispersive, that is, their different wavelength components travel at different velocities. Dispersion arises because of the velocity stratification of the Earth's interior, longer wavelengths penetrating to greater depths and hence sampling higher velocities. As a result, surface wave dispersion studies provide an important method of determining the velocity structure and seismic attenuation characteristics of the upper 600 km of the Earth.

2.1.4 Earthquake location

Earthquakes are detected by seismographs, instruments that respond to very small ground displacements, velocities, or accelerations associated with the passage of seismic waves. Since 1961 there has been an extensive and standardized global network of seismograph stations to monitor earthquake activity. The original World-Wide Standardized Seismograph Network (WWSSN), based on analogue instruments, has gradually been superseded since 1986 by the Global (Digital) Seismograph Network (GSN). By 2004 there were 136 well-distributed GSN stations worldwide, including one on the sea floor between Hawaii and California. It is hoped that this will be the first of several in oceanic areas devoid of oceanic islands for land-based stations. Digital equipment greatly facilitates processing of the data and also has the advantage that it records over a much greater dynamic range and frequency bandwidth than the earlier paper and optical recording. This is achieved by a combination of high frequency, low gain and very broadband seismometers (Butler *et al.*, 2004). Most countries have at least one GSN station and many countries also have national seismometer arrays. Together these stations not only provide the raw data for all global and regional seismological studies but also serve an important function in relation to monitoring the nuclear test ban treaty, and volcano and tsunami warning systems.

Earthquakes occurring at large, or *teleseismic*, distances from a seismograph are located by the identification of various *phases*, or seismic arrivals, on the seismograph records. Since, for example, the direct P and S waves travel at different velocities, the time separation between the arrival of the P phase and the S phase becomes progressively longer as the length of the travel path increases. By making use of a standard model for the velocity stratification of the Earth, and employing many seismic phases corresponding to different travel paths along which the seismic waves are refracted or reflected at velocity discontinuities, it is possible to translate the differences in their travel times into the distance of the earthquake from the observatory. Triangulation using distances computed in this way from many observatories then allows the location of the epicenter to be determined.

The focal depths of teleseismic events are determined by measuring the arrival time difference between the direct phase P and the phase pP (Båth, 1979). The pP phase is a short path multiple event which follows a similar path to P after first undergoing a reflection at the surface of the Earth above the focus, and so the P–pP time difference is a measure of focal depth. This method is least accurate for foci at depths of less than

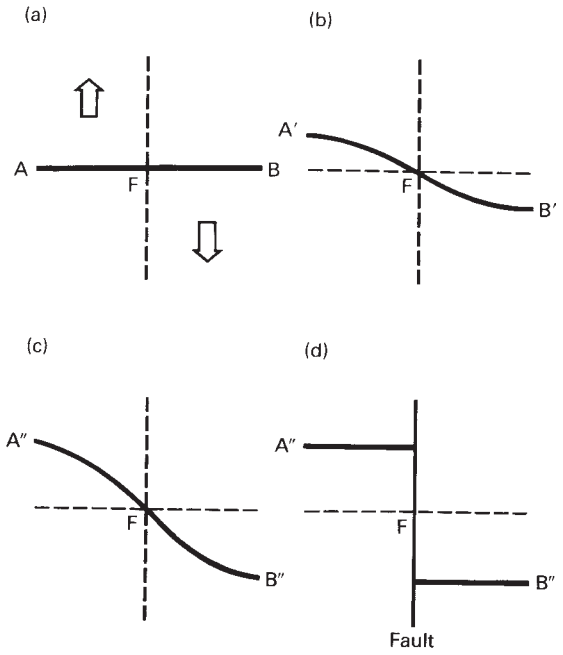


Figure 2.3 *Elastic rebound mechanism of earthquake generation.*

100 km as the P–pP time separation becomes very small. The focal depths of local earthquakes can be determined if a network of seismographs exists in the vicinity of the epicenter. In this case the focal depth is determined by triangulation in the vertical plane, using the P–S time difference to calculate the distance to the focus.

2.1.5 Mechanism of earthquakes

Most earthquakes are believed to occur according to the *elastic rebound theory*, which was developed after the San Francisco earthquake of 1906. In this theory an earthquake represents a sudden release of strain energy that has built up over a period of time.

In Fig. 2.3a a block of rock traversed by a pre-existing fracture (or fault) is being strained in such a way as eventually to cause relative motion along the plane of the fault. The line AB is a marker indicating the state of strain of the system, and the broken line the location of the fault. Relatively small amounts of strain can be accommodated by the rock (Fig. 2.3b). Eventually, however, the strain reaches the level at which it exceeds the frictional and cementing forces opposing movement along the fault plane (Fig. 2.3c). At this point fault movement occurs instantaneously (Fig. 2.3d). The 1906 San Francisco earthquake resulted from a displacement of 6.8 m along the San Andreas Fault. In this model, faulting reduces the strain in the system virtually to zero, but if the shearing forces persist, strain would again build up to the point at which fault movement occurs. The elastic rebound theory consequently implies that earthquake activity represents a stepwise response to persistent strain.

2.1.6 Focal mechanism solutions of earthquakes

The seismic waves generated by earthquakes, when recorded at seismograph stations around the world, can be used to determine the nature of the faulting associated with the earthquake, to infer the orientation of the fault plane and to gain information on the state of stress of the lithosphere. The result of such an analysis is referred to as a *focal mechanism solution* or *fault plane solution*. The technique represents a very powerful method of analyzing movements of the lithosphere, in particular those associated with plate tectonics. Information is available on a global scale as most earthquakes with a magnitude in excess of 5.5 can provide solutions, and it is not necessary to have recorders in the immediate vicinity of the earthquake, so that data are provided from regions that may be inaccessible for direct study.

According to the elastic rebound theory, the strain energy released by an earthquake is transmitted by the seismic waves that radiate from the focus. Consider the fault plane shown in Fig. 2.4 and the plane orthogonal to it, the *auxiliary plane*. The first seismic waves to arrive at recorders around the earthquake are P waves, which cause compression/dilation of the rocks through which they travel. The shaded quadrants, defined by the fault and auxiliary planes, are compressed by movement along the fault and so the first motion of the P wave arriving in these quadrants corresponds to a compression. Conversely, the unshaded quadrants are stretched or dilated by the fault movement. The first motion of the P waves in these quadrants is thus dilational. The region around the earthquake is therefore divided into four quadrants on the basis of the P wave first motions,

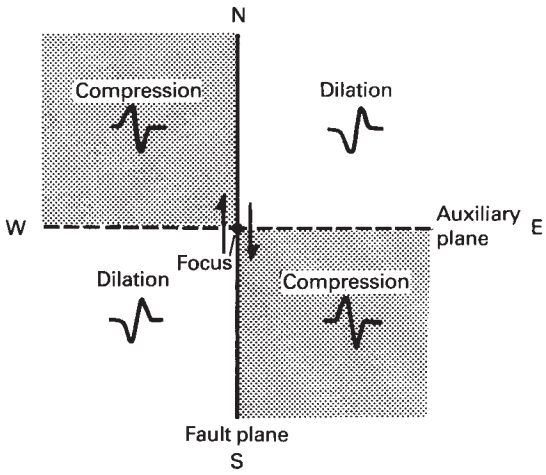


Figure 2.4 *Quadrantal distribution of compressional and dilational P wave first motions about an earthquake.*

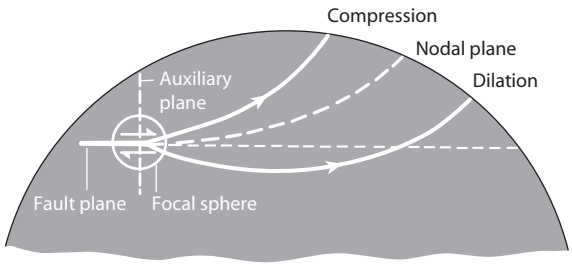


Figure 2.5 *Distribution of compressional and dilational first arrivals from an earthquake on the surface of a spherical Earth in which seismic velocity increases with depth.*

defined by the fault plane and the auxiliary plane. No P waves propagate along these planes as movement of the fault imparts only shearing motions in their directions; they are consequently known as *nodal planes*.

Simplistically, then, a focal mechanism solution could be obtained by recording an earthquake at a number of seismographs distributed around its epicenter, determining the nature of the first motions of the P waves, and then selecting the two orthogonal planes which best divide compressional from dilational first arrivals, that is, the nodal planes. In practice, however, the technique is complicated by the spheroidal shape of the Earth and the progressive increase of seismic velocity with depth that causes the seismic waves to follow curved travel paths between the focus and recorders. Consider Fig. 2.5. The dotted line represents the continuation of the fault plane, and its intersection with the Earth's surface would represent the line separating compressional and dilational first motions if the waves generated by the earthquake followed straight-line paths. The actual travel paths, however, are curved and the surface intersection of the dashed line, corresponding to the path that would have been followed by a wave leaving the focus in the direction of the fault plane, represents the actual nodal plane.

It is clear then, that simple mapping of compressional and dilational first motions on the Earth's surface cannot readily provide the focal mechanism solution. However, the complications can be overcome by considering the directions in which the seismic waves left the focal region, as it is apparent that compressions and dilations are restricted to certain angular ranges.

A focal mechanism solution is obtained firstly by determining the location of the focus by the method outlined in Section 2.1.4. Then, for each station recording the earthquake, a model for the velocity structure of the Earth is used to compute the travel path of the seismic wave from the focus to the station, and hence to calculate the direction in which the wave left the focal region. These directions are then plotted, using an appropriate symbol for compressional or dilational first motion, on an equal area projection of the lower half of the *focal sphere*, that is, an imaginary sphere of small but arbitrary radius centered on the focus (Fig. 2.5). An equal area net, which facilitates such a plot, is illustrated in Fig. 2.6. The scale around the circumference of such a net refers to the azimuth, or horizontal component of direction, while dips are plotted on the radial scale from 0° at the perimeter to 90° at the center. Planes through the focus are represented on such plots by great circles with a curvature appropriate to their dip; hence a diameter represents a vertical plane.

Let us assume that, for a particular earthquake, the fault motion is strike-slip along a near vertical fault plane. This plane and the auxiliary plane plot as orthogonal great circles on the projection of the focal sphere, as shown on Fig. 2.7. The lineation defined by the intersection of these planes is almost vertical, so it is apparent that the direction of movement along the fault is orthogonal to this intersection, that is, near horizontal. The two shaded and two unshaded regions of the projection defined by the nodal planes now correspond to the directions in which compressional and dilational

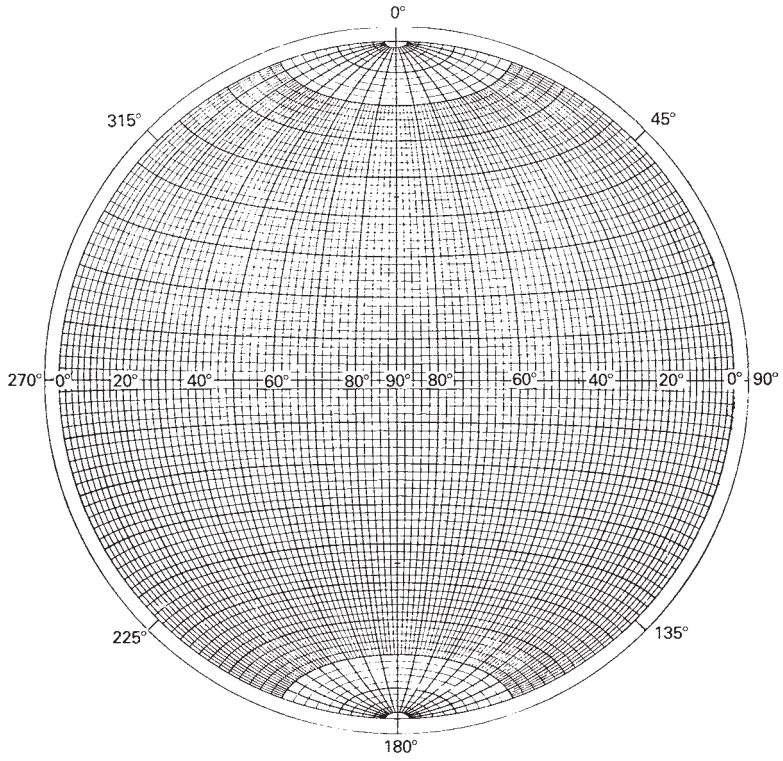


Figure 2.6 *Lambert equal area net.*

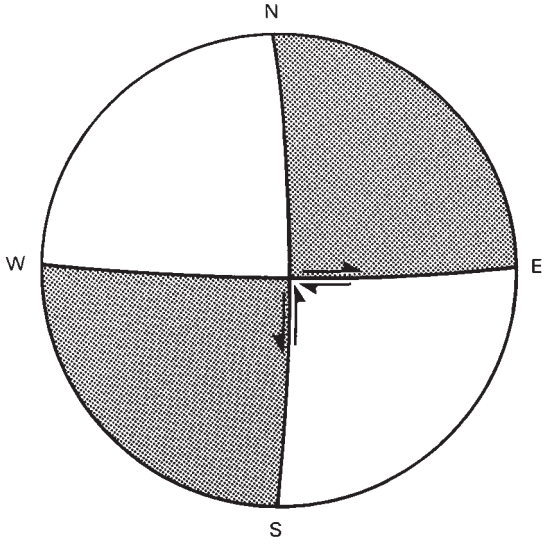


Figure 2.7 *Ambiguity in the focal mechanism solution of a strike-slip fault. Regions of compressional first motions are shaded.*

first motions, respectively, left the focal region. A focal mechanism solution is thus obtained by plotting all the observational data on the projection of the focal sphere and then fitting a pair of orthogonal planes which best divide the area of the projection into zones of compressional and dilational first motions. The more stations recording the earthquake, the more closely defined will be the nodal planes.

2.1.7 Ambiguity in focal mechanism solutions

It is apparent from Fig. 2.7 that the same distribution of compressional and dilational quadrants would be obtained if either nodal plane represented the actual fault plane. Thus, the same pattern of first motions would be obtained for sinistral motion along a north–south plane as for dextral motion along an east–west plane.

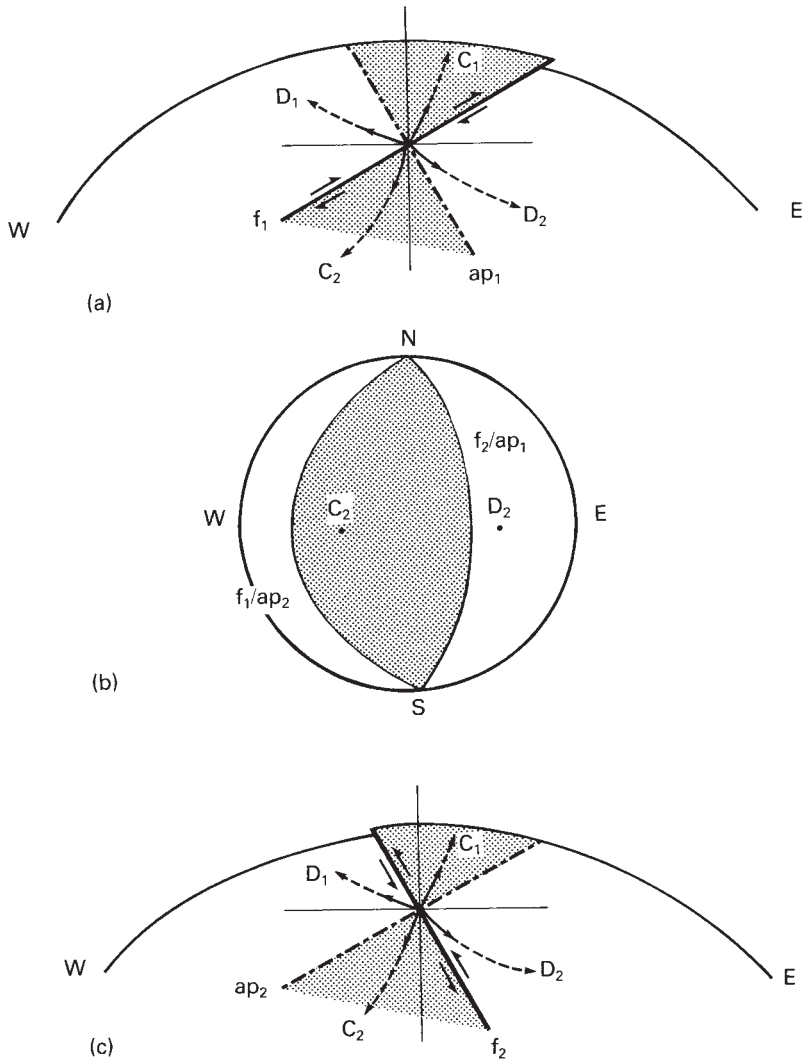


Figure 2.8 *Ambiguity in the focal mechanism solution of a thrust fault. Shaded areas represent regions of compressional first motions (C), unshaded areas represent regions of dilational first motions (D), f refers to a fault plane, ap to an auxiliary plane. Changing the nature of the nodal planes as in (a) and (c) does not alter the pattern of first motions shown in (b), the projection of the lower hemisphere of the focal sphere.*

In Fig. 2.8a an earthquake has occurred as a result of faulting along a westerly dipping thrust plane f_1. f_1 and its associated auxiliary plane ap_1 divide the region around the focus into quadrants which experience either compression or dilation as a result of the fault movement. The directions in which compressional first motions C_1 and C_2 and dilational first motions D_1 and D_2 leave the focus are shown, and C_2 and D_2 are plotted on the projection of the focal sphere in Fig. 2.8b, on which the two nodal planes are also shown. Because Fig. 2.8a is a vertical section, the first motions indicated plot along an east–west azimuth. Arrivals at stations at other azimuths would occupy other locations within the projection space. Consider now Fig. 2.8c, in which plane ap_1 becomes the fault plane f_2 and f_1 the auxiliary plane ap_2. By considering the movement along the thrust plane it is obvious that the same regions around the fault are compressed or dilated, so that an identical

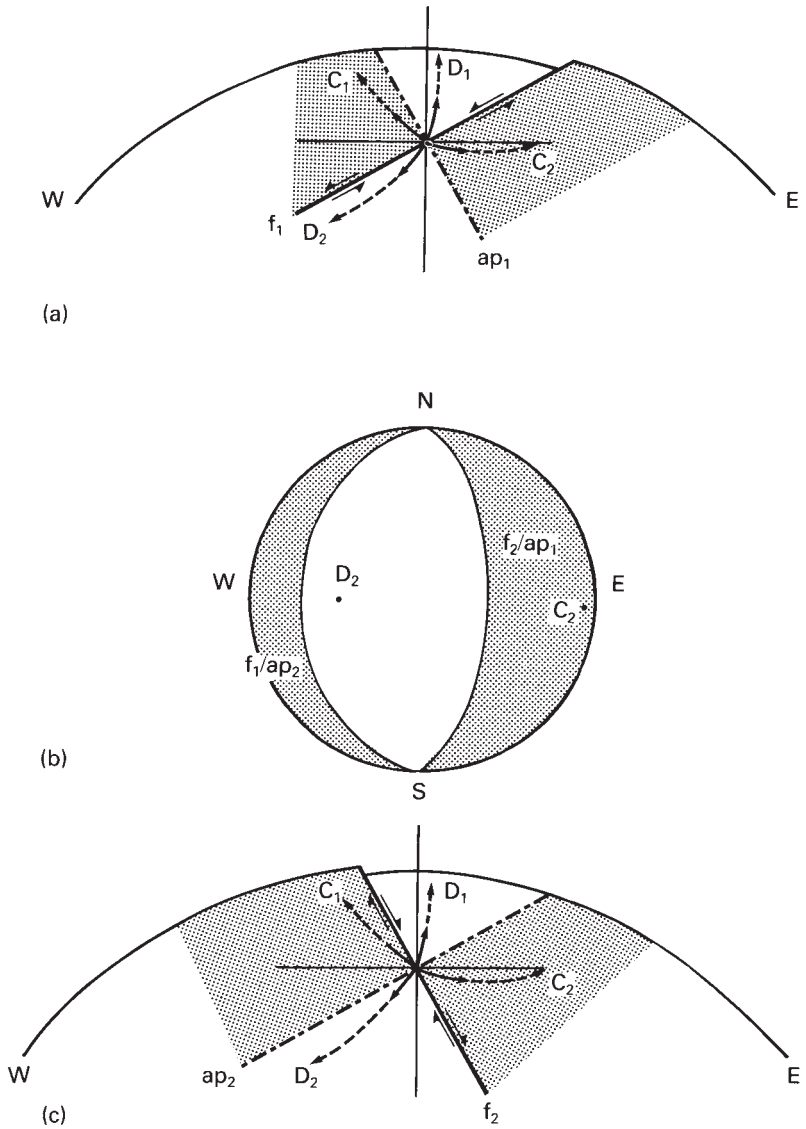


Figure 2.9 *Ambiguity in the focal mechanism solution of a normal fault. Legend as for Fig. 2.8.*

focal sphere projection is obtained. Similar results are obtained when the faulting is normal (Fig. 2.9). In theory the fault plane can be distinguished by making use of Anderson's simple theory of faulting (Section 2.10.2) which predicts that normal faults have dips of more than 45° and thrusts less than 45°. Thus, f_1 is the fault plane in Fig. 2.8 and f_2 the fault plane in Fig. 2.9.

It is apparent that the different types of faulting can be identified in a focal mechanism solution by the distinctive pattern of compressional and dilational regions on the resulting focal sphere. Indeed, it is also possible to differentiate earthquakes that have originated by a combination of fault types, such as dip-slip accompanied by some strike-slip movement. The precision with which the directions of the nodal planes can be determined is

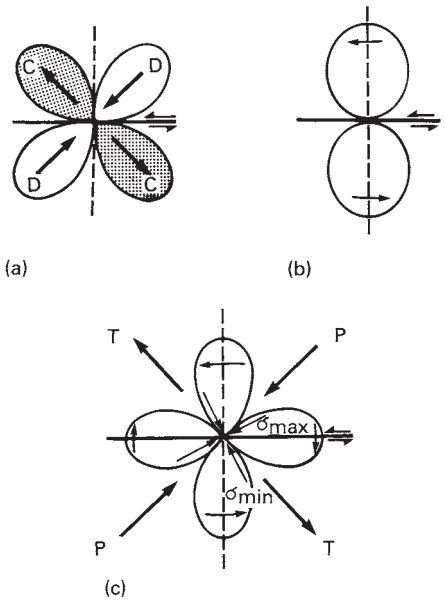


Figure 2.10 *(a) P wave radiation pattern for a type I and type II earthquake source mechanism; (b) S wave radiation pattern from a type I source (single couple); (c) S wave radiation pattern from a type II source (double couple).*

dependent upon the number and distribution of stations recording arrivals from the event. It is not possible, however, to distinguish the fault and auxiliary planes.

At one time it was believed that distinction between the nodal planes could be made on the basis of the pattern of S wave arrivals. P waves radiate into all four quadrants of the source region as shown in Fig. 2.10a. However, for this simple model, which is known as a type I, or single-couple source, S waves, whose corresponding ground motion is shearing, should be restricted to the region of the auxiliary plane (Fig. 2.10b). Recording of the S wave radiation pattern should then make it possible to determine the actual fault plane. It was found, however, that instead of this simple pattern, most earthquakes produce S wave radiation along the direction of both nodal planes (Fig. 2.10c). This observation initially cast into doubt the validity of the elastic rebound theory. It is now realized, however, that faulting occurs at an angle, typically rather less than 45% to the maximum compressive stress, σ_1, and the bisectors of the dilational and compressional quadrants, termed P and T, respectively, approximate to the directions of maximum and minimum principal compressive stress, thus giving an indication of the stress field giving rise to the earthquake (Fig. 2.10c) (Section 2.10.2).

This type II, or double-couple source mechanism gives rise to a four-lobed S wave radiation pattern (Fig. 2.10c) which cannot be used to resolve the ambiguity of a focal mechanism solution. Generally, the only constraint on the identity of the fault plane comes from a consideration of the local geology in the region of the earthquake.

2.1.8 Seismic tomography

Tomography is a technique whereby three-dimensional images are derived from the processing of the integrated properties of the medium that rays encounter along their paths through it. Tomography is perhaps best known in its medical applications, in which images of specific plane sections of the body are obtained using X-rays. Seismic tomography refers to the derivation of the three-dimensional velocity structure of the Earth from seismic waves. It is considerably more complex than medical tomography in that the natural sources of seismic waves (earthquakes) are of uncertain location, the propagation paths of the waves are unknown, and the receivers (seismographs) are of restricted distribution. These difficulties can be overcome, however, and since the late 1970s seismic tomography has provided important new information on Earth structure. The method was first described by Aki *et al.* (1977) and has been reviewed by Dziewonski & Anderson (1984), Thurber & Aki (1987), and Romanowicz (2003).

Seismic tomography makes use of the accurately recorded travel times of seismic waves from geographically distributed earthquakes at a distributed suite of seismograph stations. The many different travel paths from earthquakes to receivers cross each other many times. If there are any regions of anomalous seismic velocity in the space traversed by the rays, the travel times of the waves crossing this region are affected. The simultaneous interpretation of travel time anomalies for the many criss-crossing paths then allows the anomalous regions to be delineated, providing a three-dimensional model of the velocity space.

Both body waves and surface waves (Section 2.1.3) can be used in tomography analysis. With body waves, the actual travel times of P or S phases are utilized. The procedure with surface waves is more complex, however, as they are dispersive; that is, their velocity

depends upon their wavelength. The depth of penetration of surface waves is also wavelength-dependent, with the longer wavelengths reaching greater depths. Since seismic velocity generally increases with depth, the longer wavelengths travel more rapidly. Thus, when surface waves are utilized, it is necessary to measure the phase or group velocities of their different component wavelengths. Because of their low frequency, surface waves provide less resolution than body waves. However, they sample the Earth in a different fashion and, since either Rayleigh or Love waves (Section 2.1.3) may be used, additional constraints on shear velocity and its anisotropy are provided.

The normal procedure in seismic tomography is to assume an initial "one-dimensional" model of the velocity space in which the velocity is radially symmetrical. The travel time of a body wave from earthquake to seismograph is then equal to the sum of the travel times through the individual elements of the model. Any lateral velocity variations within the model are then reflected in variations in arrival times with respect to the mean arrival time of undisturbed events. Similarly, the dispersion of surface waves across a heterogeneous model differs from the mean dispersion through a radially symmetrical model. The method makes use of a simplifying assumption based on Fermat's Principle, which assumes that the ray paths for a radially symmetrical and laterally variable velocity model are identical if the heterogeneities are small and that the differences in travel times are caused solely by heterogeneity in the velocity structure of the travel path. This obviates the necessity of computing the new travel path implied by refractions at the velocity perturbations.

There are two main approaches to seismic tomography depending upon how the velocity heterogeneity of the model is represented. *Local methods* make use of body waves and subdivide the model space into a series of discrete elements so that it has the form of a three-dimensional ensemble of blocks. A set of linear equations is then derived which link the anomalies in arrival times to velocity variations over the different travel paths. A solution of the equations can then be obtained, commonly using matrix inversion techniques, to obtain the velocity anomaly in each block. *Global methods* express the velocity variations of the model in terms of some linear combination of continuous basic functions, such as spherical harmonic functions.

Local methods can make use of either teleseismic or local events. In the teleseismic method (Fig. 2.11) a large set of distant seismic events is recorded at a

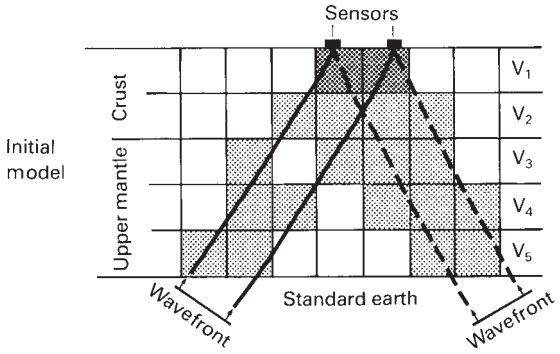


Figure 2.11 *Geometry of the teleseismic inversion method. Velocity anomalies within the compartments are derived from relative arrival time anomalies of teleseismic events (redrawn from Aki* et al.*, 1977, by permission of the American Geophysical Union. Copyright © 1977 American Geophysical Union).*

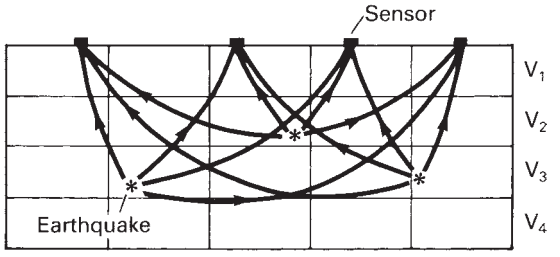


Figure 2.12 *Geometry of the local inversion method.*

network of seismographs over the volume of interest. Because of their long travel path, the incident wave fronts can be considered planar. It is assumed that deviations from expected arrival times are caused by velocity variations beneath the network. In practice, deviations from the mean travel times are computed to compensate for any extraneous effects experienced by the waves outside the volume of interest. Inversion of the series of equations of relative travel time through the volume then provides the relative velocity perturbations in each block of the model. The method can be extended by the use of a worldwide distribution of recorded teleseismic events to model the whole mantle. In the local method the seismic sources are located within the volume of interest (Fig. 2.12). In this case the location and time of the earthquakes must be accurately known, and ray-tracing methods used to construct the travel paths of the rays. The inversion

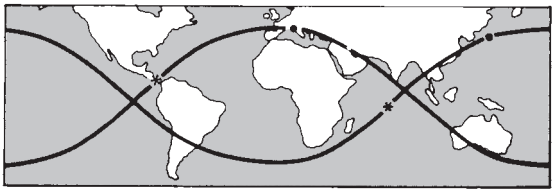


Figure 2.13 *Great circle paths from two earthquakes (stars) to recording stations (dots) (after Thurber & Aki, 1987).*

procedure is then similar to that for teleseisms. One of the uses of the resulting three-dimensional velocity distributions is to improve focal depth determinations.

Global methods commonly make use of both surface and body waves with long travel paths. If the Earth were spherically symmetrical, these surface waves would follow great circle routes. However, again making use of Fermat's Principle, it is assumed that ray paths in a heterogeneous Earth are similarly great circles, with anomalous travel times resulting from the heterogeneity. In the single-station configuration, the surface wave dispersion is measured for the rays traveling directly from earthquake to receiver. Information from only moderate-size events can be utilized, but the source parameters have to be well known. The great circle method uses multiple circuit waves, that is, waves that have traveled directly from source to receiver and have then circumnavigated the Earth to be recorded again (Fig. 2.13). Here the differential dispersion between the first and second passes is measured, eliminating any undesirable source effects. This method is appropriate to global modeling, but can only use those large magnitude events that give observable multiple circuits.

2.2 VELOCITY STRUCTURE OF THE EARTH

Knowledge of the internal layering of the Earth has been largely derived using the techniques of earthquake seismology. The shallower layers have been studied using local arrays of recorders, while the deeper layers have been investigated using global networks to detect seismic signals that have traversed the interior of the Earth.

The continental crust was discovered by Andrija Mohorovičić from studies of the seismic waves generated by the Croatia earthquake of 1909 (Fig. 2.14). Within a range of about 200 km from the epicenter, the first seismic arrivals were P waves that traveled directly from the focus to the recorders with a velocity of 5.6 km s^{-1}. This seismic phase was termed P_g. At greater ranges, however, P waves with the much higher velocity of 7.9 km s^{-1} became the first arrivals, termed the P_n phase. These data were interpreted by the standard techniques of refraction seismology, with P_n representing seismic waves that had been critically refracted at a velocity discontinuity at a depth of some 54 km. This discontinuity was subsequently named the Mohorovičić discontinuity, or Moho, and it marks the boundary between the crust and mantle. Subsequent work has demonstrated that the Moho is universally present beneath continents and marks an abrupt increase in seismic velocity to about 8 km s^{-1}. Its geometry and reflective character are highly diverse and may include one or more sub-horizontal or dipping reflectors (Cook, 2002). Continental crust is, on average, some 40 km thick, but thins to less than 20 km beneath some tectonically active rifts (e.g. Sections 7.3, 7.8.1) and thickens to up to 80 km beneath young orogenic belts (e.g. Sections 10.2.4, 10.4.5) (Christensen & Mooney, 1995; Mooney et al., 1998).

A discontinuity within the continental crust was discovered by Conrad in 1925, using similar methods. As well as the phases P_g and P_n he noted the presence of an additional phase P* (Fig. 2.15) which he interpreted as the critically refracted arrival from an interface where the velocity increased from about 5.6 to 6.3 km s^{-1}. This interface was subsequently named the Conrad discontinuity. Conrad's model was readily adopted by early petrologists who believed that two layers were necessarily present in the continental crust. The upper layer, rich in silicon and aluminum, was called the SIAL and was believed to be the source of granitic magmas, while the lower, silicon- and magnesium-rich layer or SIMA was believed to be the source of basaltic magmas. It is now known, however, that the upper crust has a composition more mafic than granite (Section 2.4.1), and that the majority of basaltic magmas originate in the mantle. Consequently, the petrological necessity of a two-layered crust no

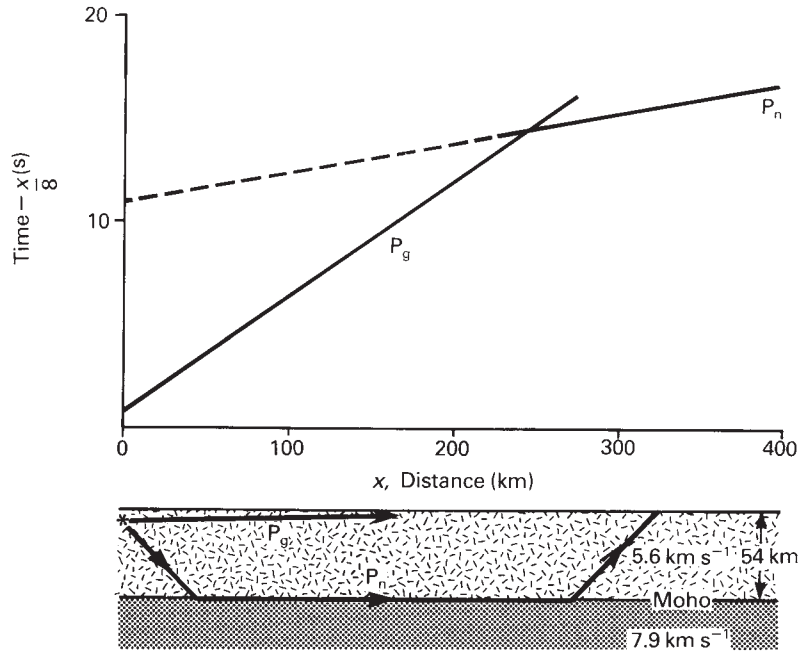


Figure 2.14 *Reduced time–distance relationship for direct waves (P_g) and waves critically refracted at the Moho (P_n) from an earthquake source.*

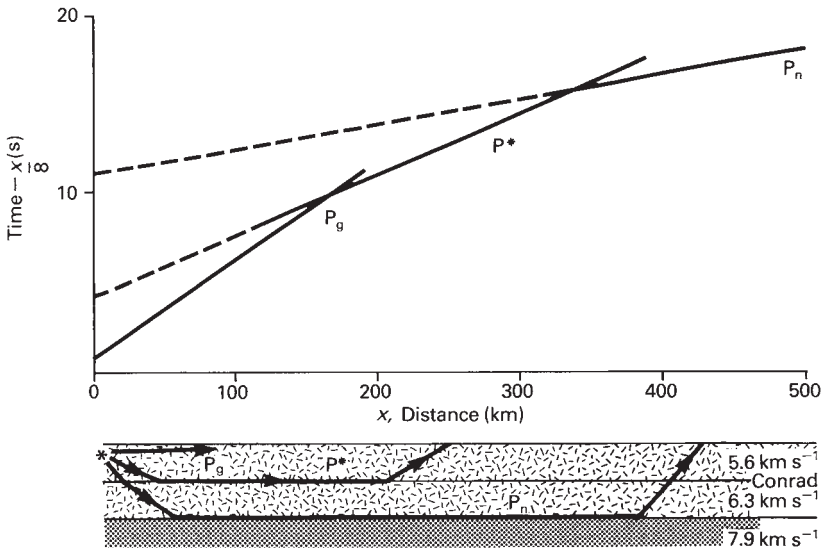


Figure 2.15 *Reduced time–distance relationship for direct waves (P_g), waves critically refracted at the Conrad discontinuity (P*) and waves critically refracted at the Moho (P_n) from an earthquake source.*

longer exists and, where applicable, it is preferable to use the terms upper and lower crust. Unlike the Moho, the Conrad discontinuity is not always present within the continental crust, although the seismic velocity generally increases with depth.

In some regions the velocity structure of continental crust suggests a natural division into three layers. The velocity range of the middle crustal layer generally is taken to be 6.4–6.7 km s^{-1}. The typical velocity range of the lower crust, where a middle crust is present, is 6.8–7.7 km s^{-1} (Mooney *et al.*, 1998). Examples of the velocity structure of continental crust in a tectonically active rift, a rifted margin, and a young orogenic belt are shown in Figs 7.5, 7.32a, and 10.7, respectively.

The oceanic crust has principally been studied by explosion seismology. The Moho is always present and the thickness of much of the oceanic crust is remarkably constant at about 7 km irrespective of the depth of water above it. The internal layering of oceanic crust and its constancy over very wide areas will be discussed later (Section 2.4.4).

In studying the deeper layering of the Earth, seismic waves with much longer travel paths are employed. The velocity structure has been built up by recording the travel times of body waves over the full range of possible epicentral angles. By assuming that the Earth is radially symmetrical, it is possible to invert the travel time data to provide a model of the velocity structure. A modern determination of the velocity–depth curve (Kennett *et al.*, 1995) for both P and S waves is shown in Fig. 2.16.

Velocities increase abruptly at the Moho in both continental and oceanic environments. A low velocity zone (LVZ) is present between about 100 and 300 km depth, although the depth to the upper boundary is very variable (Section 2.12). The LVZ appears to be universally present for S waves, but may be absent in certain regions for P waves, especially beneath ancient shield areas. Between 410 and 660 km velocity increases rapidly in a stepwise fashion within the mantle transition zone that separates the upper mantle from the lower mantle. Each velocity increment probably corresponds to a mineral phase change to a denser form at depth (Section 2.8.5). Both P and S velocities increase progressively in the lower mantle.

The Gutenberg discontinuity marks the core–mantle boundary at a depth of 2891 km, at which the velocity of P waves decreases abruptly. S waves are not transmitted through the outer core, which is consequently

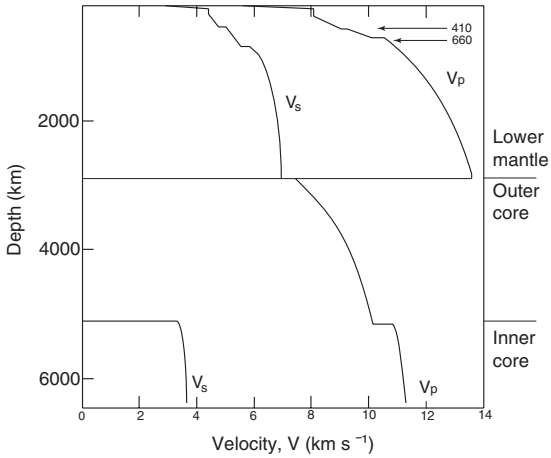


Figure 2.16 *Seismic wave velocities as a function of depth in the Earth showing the major discontinuities. AK 135 Earth model specified by Kennett* et al.*, 1995 (after Helffrich & Wood, 2001, with permission from* Nature ***412****, 501–7. Copyright © 2001 Macmillan Publishers Ltd.).*

believed to be in a fluid state. The geomagnetic field (Section 3.6.4) is believed to originate by the circulation of a good electrical conductor in this region. At a depth of 5150 km the P velocity increases abruptly and S waves are once again transmitted. This inner core is thus believed to be solid as a result of the enormous confining pressure. There appears to be no transition zone between inner and outer core, as was originally believed.

2.3 COMPOSITION OF THE EARTH

All bodies in the solar system are believed to have been formed by the condensation and accretion of the primitive interstellar material that made up the solar nebula. The composition of the Sun is the same as the average composition of this material. Gravitational energy was released during accretion, and together with the radioactive decay of short-lived radioactive nuclides eventually led to heating of the proto-Earth so that it differentiated into a radially symmetric body made up of a series of shells whose density increased towards its

center. The differentiation prevents any estimate being made of the overall composition of the Earth by direct sampling. However, it is believed that meteorites are representatives of material within the solar nebula and that estimates of the Earth's composition can be made from them. The presence of metallic and silicate phases in meteorites is taken to indicate that the Earth consists of an iron/nickel core surrounded by a lower density silicate mantle and crust.

Seismic data, combined with knowledge of the mass and moment of inertia of the Earth, have revealed that the mean atomic weight of the Earth is about 27, with a contribution of 22.4 from the mantle and crust and 47.0 from the core. No single type of meteorite possesses an atomic weight of 27, the various types of chondrite being somewhat lower and iron meteorites considerably higher. However, it is possible to mix the proportions of different meteorite compositions in such a way as to give both the correct atomic weight and core/mantle ratio. Three such models are given in Table 2.1.

It is apparent that at least 90% of the Earth is made up of iron, silicon, magnesium, and oxygen, with the bulk of the remainder comprising calcium, aluminum, nickel, sodium, and possibly sulfur.

Table 2.1 *Estimates of the bulk composition of the Earth and Moon (in weight percent) (from Condie, 1982a).*

	Earth			Moon
	1	2	3	4
Fe	34.6	29.3	29.9	9.3
O	29.5	30.7	30.9	42.0
Si	15.2	14.7	17.4	19.6
Mg	12.7	15.8	15.9	18.7
Ca	1.1	1.5	1.9	4.3
Al	1.1	1.3	1.4	4.2
Ni	2.4	1.7	1.7	0.6
Na	0.6	0.3	0.9	0.07
S	1.9	4.7	–	0.3

1: 32.4% iron meteorite (with 5.3% FeS) and 67.6% oxide portion of bronzite chondrites.
2: 40% type I carbonaceous chondrite, 50% ordinary chondrite, and 10% iron meteorite (containing 15% sulfur).
3: Nonvolatile portion of type I carbonaceous chondrites with FeO/FeO + MgO of 0.12 and sufficient SiO_2 reduced to Si to yield a metal/silicate ratio of 32/68.
4: Based on Ca, Al, Ti = 5×type I carbonaceous chondrites, FeO = 12% to accommodate lunar density, and Si/Mg = chondritic ratio.

2.4 THE CRUST

2.4.1 The continental crust

Only the uppermost part of the crust is available for direct sampling at the surface or from boreholes. At greater depths within the crust, virtually all information about its composition and structure is indirect. Geologic studies of high grade metamorphic rocks that once resided at depths of 20–50 km and have been brought to the surface by subsequent tectonic activity provide some useful information (Miller & Paterson, 2001a; Clarke *et al.*, 2005). Foreign rock fragments, or *xenoliths*, that are carried from great depths to the Earth's surface by fast-rising magmas (Rudnick, 1992) also provide samples of deep crustal material. In addition, much information about the crust has been derived from knowledge of the variation of seismic velocities with depth and how these correspond to experimental determinations of velocities measured over ranges of temperature and pressure consistent with crustal conditions. Pressure increases with depth at a rate of about 30 MPa km^{-1}, mainly due to the lithostatic confining pressure of the overlying rocks, but also, in some regions, with a contribution from tectonic forces. Temperature increases at an average rate of about 25 °C km^{-1}, but decreases to about half this value at the Moho because of the presence of radioactive heat sources within the crust (Section 2.13). Collectively, the observations from both geologic and geophysical studies show that the continental crust is vertically stratified in terms of its chemical composition (Rudnick & Gao, 2003).

The variation of seismic velocities with depth (Section 2.2) results from a number of factors. The increase of pressure with depth causes a rapid increase in incompressibility, rigidity, and density over the topmost 5 km as pores and fractures are closed. Thereafter the increase of these parameters with pressure is balanced by the decrease resulting from thermal expansion with increasing temperature so that there is little further change in velocity with depth. Velocities change with chemical composition, and also with changes in mineralogy resulting from phase changes. Abrupt velocity discontinuities are usually caused by

changes in chemical composition, while more gradational velocity boundaries are normally associated with phase changes that occur over a discrete vertical interval.

Models for the bulk chemical composition of the continental crust vary widely because of the difficulty of making such estimates. McLennan & Taylor (1996) pointed out that the flow of heat from the continental crust (Section 2.13) provides a constraint on the abundance of the heat producing elements, K, Th, and U, within it, and hence on the silica content of the crust. On this basis they argue that on average the continental crust has an andesitic or granodioritic composition with K_2O no more than 1.5% by weight. This is less silicic than most previous estimates. The abundance of the heat producing elements, and other "incompatible" elements, in the continental crust is of great importance because the degree to which they are enriched in the crust reflects the extent to which they are depleted in the mantle.

2.4.2 Upper continental crust

Past theories of crustal construction suggested that the upper continental crust was made up of rocks of granitic composition. That this is not the case is evident from the widespread occurrence of large negative gravity anomalies over granite plutons. These anomalies demonstrate that the density of the plutons (about 2.67 $Mg\,m^{-3}$) is some 0.10–0.15 $Mg\,m^{-3}$ lower than the average value of the upper crust. The mean composition of the upper crust can be estimated, albeit with some uncertainty due to biasing, by determining the mean composition of a large number of samples collected worldwide and from analyses of sedimentary rocks that have sampled the crust naturally by the process of erosion (Taylor & Scott, 1985; Gao *et al.*, 1998). This composition corresponds to a rock type between granodiorite and diorite, and is characterized by a relatively high concentration of the heat-producing elements.

2.4.3 Middle and lower continental crust

For a 40 km thick average global continental crust (Christensen & Mooney, 1995; Mooney *et al.*, 1998), the middle crust is some 11 km thick and ranges in depth from 12 km, at the top, to 23 km at the bottom (Rudnick & Fountain, 1995; Gao *et al.*, 1998). The average lower crust thus begins at 23 km depth and is 17 km thick. However, the depth and thickness of both middle and lower crust vary considerably from setting to setting. In tectonically active rifts and rifted margins, the middle and lower crust generally are thin. The lower crust in these settings can range from negligible to more than 10 km thick (Figs 7.5, 7.32a). In Mesozoic–Cenozoic orogenic belts where the crust is much thicker, the lower crust may be up to 25 km thick (Rudnick & Fountain, 1995).

The velocity range of the lower crust (6.8–7.7 $km\,s^{-1}$, Section 2.2) cannot be explained by a simple increase of seismic velocity with depth. Consequently, either the chemical composition must be more mafic, or denser, high-pressure phases are present. Information derived from geologic studies supports this conclusion, indicating that continental crust becomes denser and more mafic with depth. In addition, the results from these studies show that the concentration of heat-producing elements decreases rapidly from the surface downwards. This decrease is due, in part, to an increase in metamorphic grade but is also due to increasing proportions of mafic lithologies.

In areas of thin continental crust, such as in rifts and at rifted margins, the middle and lower crust may be composed of low- and moderate-grade metamorphic rocks. In regions of very thick crust, such as orogenic belts, the middle and lower crust typically are composed of high-grade metamorphic mineral assemblages. The middle crust in general may contain more evolved and less mafic compositions compared to the lower crust. Metasedimentary rocks may be present in both layers. If the lower crust is dry, its composition could correspond to a high-pressure form of granulite ranging in composition from granodiorite to diorite (Christensen & Fountain, 1975; Smithson & Brown, 1977), and containing abundant plagioclase and pyroxene minerals. In the overthickened roots of orogens, parts of the lower crust may record the transition to the eclogite facies, where plagioclase is unstable and mafic rocks transform into very dense, garnet-, pyroxene-bearing assemblages (Section 9.9). If the lower crust is wet, basaltic rocks would occur in the form of amphibolite. If mixed with more silicic material, this would have a seismic velocity in the correct range. Studies of exposed sections of ancient lower crust suggest that both dry and wet rock types typically are present (Oliver, 1982; Baldwin *et al.*, 2003).

Another indicator of lower crust composition is the elastic deformation parameter Poisson's ratio, which can be expressed in terms of the ratio of P and S wave velocities for a particular medium. This parameter varies systematically with rock composition, from approximately 0.20 to 0.35. Lower values are characteristic of rocks with high silica content, and high values with mafic rocks and relatively low silica content. For example, beneath the Main Ethiopian Rift in East Africa (Fig. 7.2) Poisson's ratios vary from 0.27 to 0.35 (Dugda *et al.*, 2005). By contrast, crust located outside the rift is characterized by varying from 0.23 to 0.28. The higher ratios beneath the rift are attributed to the intrusion and extensive modification of the lower crust by mafic magma (Fig. 7.5).

Undoubtedly, the lower crust is compositionally more complex than suggested by these simple geophysical models. Studies of deep crustal xenoliths and crustal contaminated magmas indicate that there are significant regional variations in its composition, age, and thermal history. Deep seismic reflection investigations (Jackson, H.R., 2002; van der Velden *et al.*, 2004) and geologic studies of ancient exposures (Karlstrom & Williams, 1998; Miller & Paterson, 2001a; Klepeis *et al.*, 2004) also have shown that this compositional complexity is matched by a very heterogeneous structure. This heterogeneity reflects a wide range of processes that create and modify the lower crust. These processes include the emplacement and crystallization of magma derived from the mantle, the generation and extraction of crustal melts, metamorphism, erosion, tectonic burial, and many other types of tectonic reworking (Sections 9.8, 9.9).

2.4.4 The oceanic crust

The oceanic crust (Francheteau, 1983) is in isostatic equilibrium with the continental crust according to the Airy mechanism (Section 2.11.2), and is consequently much thinner. Seismic refraction studies have confirmed this and show that oceanic crust is typically 6–7 km thick beneath an average water depth of 4.5 km. Thicker oceanic crust occurs where the magma supply rate is anomalously high due to higher than normal temperatures in the upper mantle. Conversely, thinner than normal crust forms where upper mantle temperatures are anomalously low, typically because of a very low rate of formation (Section 6.10).

Table 2.2 *Oceanic crustal structure (after Bott, 1982).*

	***P* velocity ($km\,s^{-1}$)**	**Average thickness (km)**
Water	1.5	4.5
Layer 1	1.6–2.5	0.4
Layer 2	3.4–6.2	1.4
Layer 3	6.4–7.0	5.0
	Moho	
Upper mantle	7.4–8.6	

The earliest refraction surveys produced time–distance data of relatively low accuracy that, on simple inversion using plane-layered models, indicated the presence of three principal layers. The velocities and thicknesses of these layers are shown in Table 2.2. More recent refraction studies, employing much more sophisticated equipment and interpretational procedures (Kennett B.L.N., 1977), have shown that further subdivision of the main layers is possible (Harrison & Bonatti, 1981) and that, rather than a structure in which velocities increase downwards in discrete jumps, there appears to be a progressive velocity increase with depth (Kennett & Orcutt, 1976; Spudich & Orcutt, 1980). Figure 2.17 compares the velocity structure of the oceanic crust as determined by early and more recent investigations.

2.4.5 Oceanic layer 1

Layer 1 has been extensively sampled by coring and drilling. Seabed surface materials comprise unconsolidated deposits including terrigenous sediments carried into the deep oceans by turbidity currents, and pelagic deposits such as brown zeolite clays, calcareous and silicic oozes, and manganese nodules. These deep-sea sediments are frequently redistributed by bottom currents or contour currents, which are largely controlled by thermal and haline anomalies within the oceans. The dense, cold saline water produced at the poles sinks and underflows towards equatorial regions, and is deflected by the Coriolis force. The resulting currents give rise to sedimentary deposits that are termed *contourites* (Stow & Lovell, 1979).

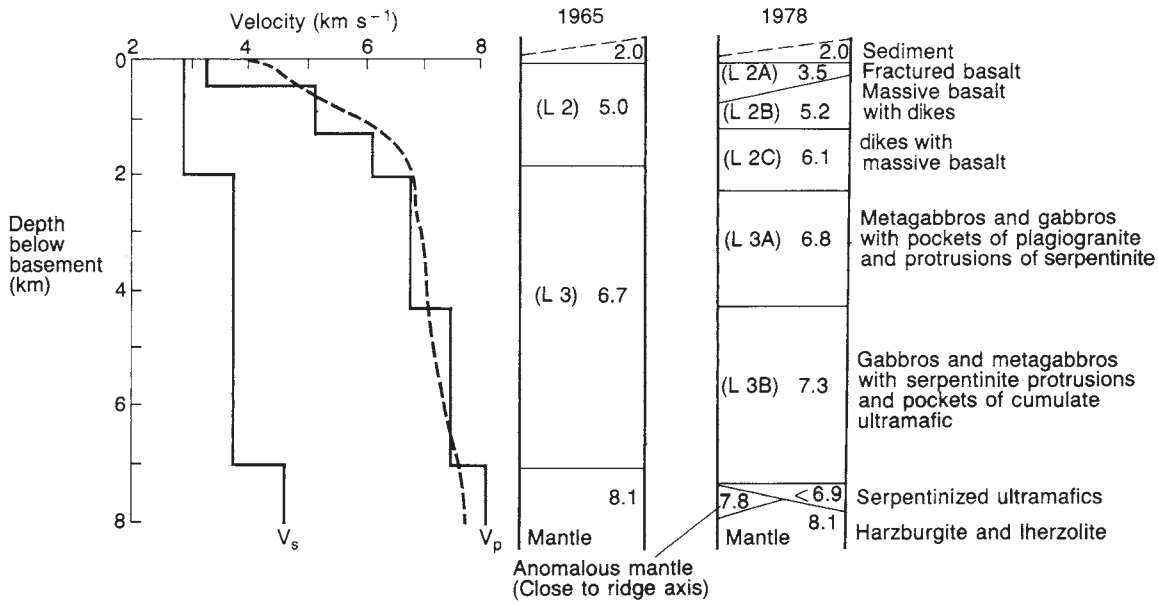


Figure 2.17 *P and S wave velocity structure of the oceanic crust and its interpretation in terms of layered models proposed in 1965 and 1978. Numbers refer to velocities in km s^{-1}. Dashed curve refers to gradational increase in velocity with depth deduced from more sophisticated inversion techniques (after Spudich & Orcutt, 1980 and Harrison & Bonatti, 1981).*

Layer 1 is on average 0.4 km thick. It progressively thickens away from the ocean ridges, where it is thin or absent. There is, however, a systematic difference in the sediment thicknesses of the Pacific and Atlantic/Indian oceans. The former is rimmed by trenches, that trap sediments of continental origin, and the latter are not, allowing greater terrestrial input. The interface between layer 1 and layer 2 is considerably more rugged than the seabed, because of the volcanic and faulted nature of layer 2. Within layer 1 are a number of horizons that show up as prominent reflectors on seismic reflection records. Edgar (1974) has described the acoustic stratigraphy in the North Atlantic, where up to four suprabasement reflectors are found (Fig. 2.18). Horizon A corresponds to an Eocene chert, although deep sea drilling indicates that it maintains its reflective character even when little or no chert is present. In such locations it may correspond to an early Cenozoic hiatus beneath the chert. Horizon A* occurs beneath A, and represents the interface between Late Cretaceous/Paleogene metal-rich clays and underlying euxinic black clays. Horizon B represents the base of the black clays, where they overlie a Late Jurassic/Lower Cretaceous limestone. Horizon B may represent a sedimentary horizon, although it has also been identified as basalt similar to that at the top of layer 2.

Reflectors similar to A and B have been identified in the Pacific and Caribbean, where they are termed A′, B′ and A″, B″, respectively.

2.4.6 Oceanic layer 2

Layer 2 is variable in its thickness, in the range 1.0–2.5 km. Its seismic velocity is similarly variable in the range 3.4–6.2 km s^{-1}. This range is attributable to either consolidated sediments or extrusive igneous material. Direct sampling and dredging of the sediment-free crests of ocean ridges, and the necessity of a highly magnetic lithology at this level (Section 4.2), overwhelmingly prove an igneous origin. The basalts recovered are olivine tholeiites containing calcic plagioclase, and are poor in potassium, sodium, and the incompatible elements (Sun *et al.*, 1979). They exhibit very little areal variation in major element composition, with the exception of locations close to oceanic islands (Section 5.4).

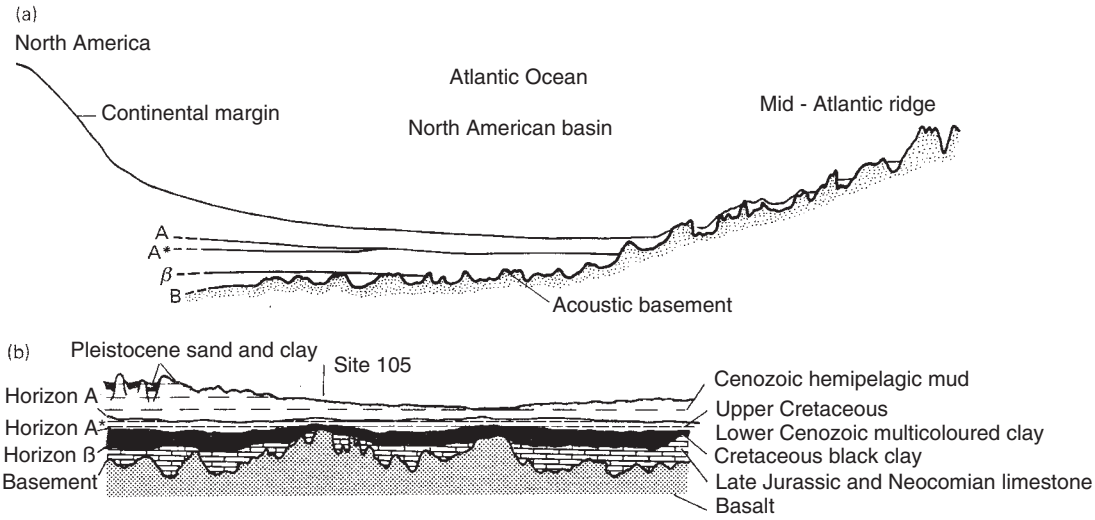


Figure 2.18 *(a) Major seismic reflectors in the western Atlantic Ocean. (b) Corresponding lithologies determined by deep sea drilling (after Edgar, 1974, Fig. 1. Copyright © 1974, with kind permission of Springer Science and Business Media).*

Three subdivisions of layer 2 have been recognized. Sublayer 2A is only present on ocean ridges near eruptive centers in areas affected by hydrothermal circulation of sea water, and ranges in thickness from zero to 1 km. Its porous, rubbly nature, as indicated by a P wave velocity of 3.6 km s^{-1}, permits such circulation. The very low velocities (2.1 km s^{-1}) of the top of very young layer 2 located on the Mid-Atlantic Ridge (Purdy, 1987) probably indicate a porosity of 30–50%, and the much higher velocities of older layer 2 imply that the porosity must be reduced quite rapidly after its formation. Sublayer 2B forms the normal acoustic basement of layer 1 when sublayer 2A is not developed. Its higher velocity of 4.8–5.5 km s^{-1} suggests a lower porosity. With time layer 2A may be converted to layer 2B by the infilling of pores by secondary minerals such as calcite, quartz, and zeolites. Sublayer 2C is about 1 km thick, where detected, and its velocity range of 5.8–6.2 km s^{-1} may indicate a high proportion of intrusive, mafic rocks. This layer grades downwards into layer 3.

The DSDP/ODP drill hole 504B, that drilled through the top 1800 m of igneous basement in 6 Ma old crust on the Costa Rica Rift, in the eastern central Pacific, encountered pillow lavas and dikes throughout. It revealed that, at least for this location, the layer 2/3 seismic boundary lies within a dike complex and is associated with gradual changes in porosity and alteration (Detrick *et al.*, 1994).

2.4.7 Oceanic layer 3

Layer 3 is the main component of the oceanic crust and represents its plutonic foundation (Fox & Stroup, 1981). Some workers have subdivided it into sublayer 3A, with a velocity range of 6.5–6.8 km s^{-1}, and a higher velocity lower sublayer 3B (7.0–7.7 km s^{-1}) (Christensen & Salisbury, 1972), although the majority of seismic data can be explained in terms of a layer with a slight positive velocity gradient (Spudich & Orcutt, 1980).

Hess (1962) suggested that layer 3 was formed from upper mantle material whose olivine had reacted with water to varying degrees to produce serpentinized peridotite, and, indeed, 20–60% serpentinization can explain the observed range of P wave velocities. However for oceanic crust of normal thickness (6–7 km) this notion can now be discounted, as the value of Poisson's ratio for layer 3A, which can be estimated directly from a knowledge of both P and S wave velocities, is much lower than would be expected for serpentinized peridotite. In fact, Poisson's ratio for layer 3A is more in accord with a gabbroic composition, which also provides seismic velocities in the observed range. It is possible,

however, that all or at least part of layer 3B, where recognized, consists of serpentinized ultramafic material.

The concept of a predominantly gabbroic layer 3 is in accord with models suggested for the origin of oceanic lithosphere (Section 6.10). These propose that layer 3 forms by the crystallization of a magma chamber or magma chambers, with an upper layer, possibly corresponding to sublayer 3A, of isotropic gabbro and a lower layer, possibly corresponding to 3B, consisting of cumulate gabbro and ultramafic rocks formed by crystal settling. This layering has been confirmed by direct observation and sampling by submersible on the Vema Fracture Zone in the North Atlantic (Auzende *et al.*, 1989).

2.5 OPHIOLITES

The study of oceanic lithosphere has been aided by investigations of characteristic rock sequences on land known as ophiolites (literally "snake rock", referring to the similarity of the color and texture to snakeskin; see Nicolas, 1989, for a full treatment of this topic). Ophiolites usually occur in collisional orogens (Section 10.4), and their association of deep-sea sediments, basalts, gabbros, and ultramafic rocks suggests that they originated as oceanic lithosphere and were subsequently thrust up into their continental setting by a process known as *obduction* (Dewey, 1976; Ben-Avraham *et al.*, 1982; Section 10.6.3). The complete ophiolite sequence (Gass, 1980) is shown in Table 2.3. The analogy of ophiolites with oceanic lithosphere is supported by the gross similarity in chemistry (although there is considerable difference in detail), metamorphic grades corresponding to temperature gradients existing under spreading centers, the presence of similar ore minerals, and the observation that the sediments were formed in deep water (Moores, 1982). Salisbury & Christensen (1978) have compared the velocity structure of the oceanic lithosphere with seismic velocities measured in samples from the Bay of Islands ophiolite complex in Newfoundland, and concluded that the determined velocity stratigraphies are identical. Figure 2.19 shows the correlation between the oceanic lithosphere and three well-studied ophiolite bodies.

At one time it seemed that investigations of the petrology and structure of the oceanic lithosphere could conveniently be accomplished by the study of

Table 2.3 *Correlation of ophiolite stratigraphy with the oceanic lithosphere (after Gass, 1980 with permission from the Ministry of Agriculture and Natural Resources, Cyprus).*

Complete ophiolite sequence		***Oceanic correlation***
Sediments		Layer 1
Mafic volcanics, commonly pillowed, merging into Mafic sheeted dike complex	}	Layer 2
High level intrusives Trondhjemites Gabbros	}	Layer 3
Layered cumulates Olivine gabbros Pyroxenites Peridotites	}	— Moho —
Harzburgite, commonly serpentinized ± lherzolite, dunite, chromitite		Upper mantle

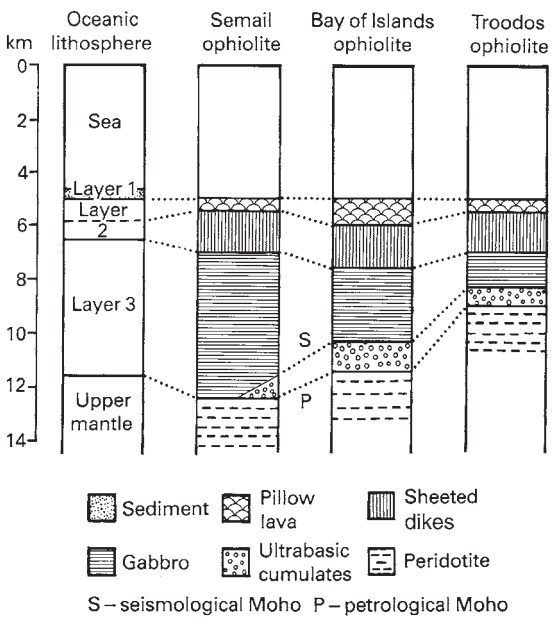


Figure 2.19 *Comparison of oceanic crustal structure with ophiolite complexes (after Mason, 1985, with permission from Blackwell Publishing).*

ophiolite sequences on land. However, this simple analogy has been challenged, and it has been suggested that ophiolites do not represent typical oceanic lithosphere, and were not emplaced exclusively during continental collision (Mason, 1985).

Dating of events indicates that obduction of many ophiolites occurred very soon after their creation. Continental collision, however, normally occurs a long time after the formation of a mid-ocean ridge, so that the age of the sea floor obducted should be considerably greater than that of the collisional orogeny. Ophiolites consequently represent lithosphere that was obducted while young and hot. Geochemical evidence (Pearce, 1980; Elthon, 1991) has suggested that the original sites of ophiolites were backarc basins (Section 9.10; Cawood & Suhr, 1992), Red Sea-type ocean basins, or the forearc region of subduction zones (Flower & Dilek, 2003). The latter setting seems at first to be an unlikely one. However, the petrology and geochemistry of the igneous basement of forearcs, which is very distinctive, is very comparable to that of many ophiolites. Formation in a forearc setting could also explain the short time interval between formation and emplacement, and the evidence for the "hot" emplacement of many ophiolites. A backarc or forearc origin is also supported by the detailed geochemistry of the lavas of most ophiolites, which indicates that they are derived from melts that formed above subduction zones.

There have been many different mechanisms proposed for ophiolite obduction, none of which can satisfactorily explain all cases. It must thus be recognized that there may be several operative mechanisms and that, although certainly formed by some type of accretionary process, ophiolite sequences may differ significantly, notably in terms of their detailed geochemistry, from lithosphere created at mid-ocean ridge crests in the major ocean basins.

Although many ophiolites are highly altered and tectonized, because of the way in which they are uplifted and emplaced in the upper crust, there are definite indications that there is more than one type of ophiolite. Some have the complete suite of units listed in Table 2.3 and illustrated in Fig. 2.19, others consist solely of deep-sea sediments, pillow lavas, and serpentinized peridotite, with or without minor amounts of gabbro. If present these gabbros often occur as intrusions within the serpentinized peridotite. These latter types are remarkably similar to the inferred nature of the thin oceanic crust that forms where magma supply rates are low. This type of crust is thought to form when the rate of formation of the crust is very low (Section 6.10), in the vicinity of transform faults at low accretion rates (Section 6.7), and in the initial stages of ocean crust formation at nonvolcanic passive continental margins (Section 7.7.2). It seems probable that Hess (1962), in suggesting that layer 3 of the oceanic crust is serpentinized mantle, was in part influenced by his experience and knowledge of ophiolites of this type in the Appalachian and Alpine mountain belts.

2.6 METAMORPHISM OF OCEANIC CRUST

Many of the rocks sampled from the ocean basins show evidence of metamorphism, including abundant greenschist facies assemblages and alkali metasomatism: In close proximity to such rocks, however, are found completely unaltered species.

It is probable that this metamorphism is accomplished by the hydrothermal circulation of seawater within the oceanic crust. There is much evidence for the existence of such circulation, such as the presence of metalliferous deposits which probably formed by the leaching and concentration of minerals by seawater, observations of active hydrothermal vents on ocean ridges (Section 6.5), and the observed metamorphism within ophiolite sequences.

Hydrothermal circulation takes place by convective flow, probably through the whole of the oceanic crust (Fyfe & Lonsdale, 1981), and is of great significance. It influences models of heat production, as it has been estimated that approximately 25% of the heat escaping from the Earth's surface is vented at the mid-ocean ridges. The circulation must modify the chemistry of the ocean crust, and consequently will affect the chemical relationship of lithosphere and asthenosphere over geologic time because of the recycling of lithosphere that occurs at subduction zones. It is also responsible for the formation of certain economically important ore deposits, particularly massive sulfides.

These hydrothermal processes are most conveniently studied in the metamorphic assemblages of ophiolite complexes, and the model described below has been derived by Elthon (1981).

Hydrothermal metamorphism of pillow lavas and other extrusives gives rise to low-temperature ($<230^{\circ}C$)

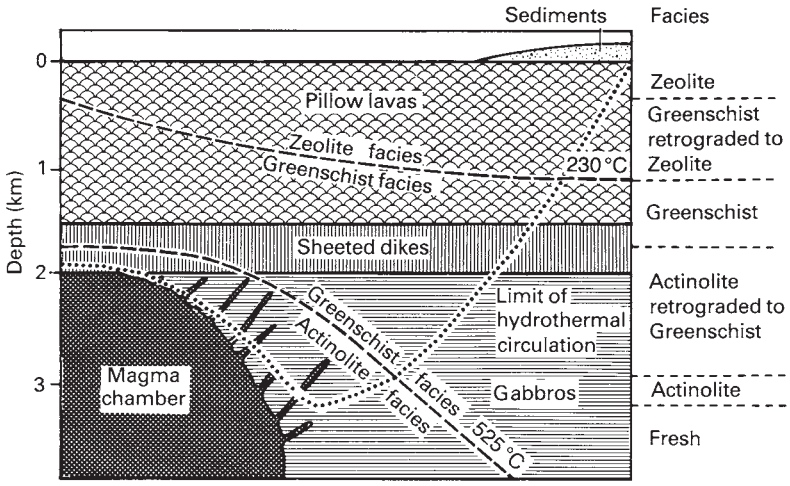


Figure 2.20 *Schematic model for hydrothermal metamorphism of the oceanic crust at a spreading center (redrawn from Elthon, 1981).*

and greenschist facies assemblages (Fig. 2.20). The distribution of alteration is highly irregular, and is controlled by the localized fissuring of the extrusive rocks. Higher temperature metamorphism is widespread within the sheeted dike complex, producing assemblages typical of the actinolite facies, although pockets of unaltered rock do occur. The highest metamorphic temperatures are achieved at the base of the sheeted dike complex and the upper part of the gabbroic section. Rarely, retrograde rocks of the greenschist facies occur at this level. Alteration decreases to only about 10% within the top kilometer of the gabbroic section and thereafter metamorphism is restricted to the locality of fissures and dikes, although metamorphism does not completely terminate at depth. According to this model, seawater circulation occurs extensively in the upper 3 km of the crust, producing the metamorphic assemblages and cooling the crust. High-temperature metamorphism only occurs near the spreading center. At depth the circulation becomes diminished as secondary minerals are deposited within the flow channels.

As the ridge spreads continuously, oceanic lithosphere is moved laterally from the heat source and undergoes retrograde metamorphism. This depends upon an adequate water supply, as water distribution is the major control of metamorphic grade. The absence of sufficient water allows the preservation of relict high temperature assemblages. The heterogeneous nature of the distribution of metamorphic facies is consequently explained by a similarly heterogeneous distribution of circulating fluids rather than extreme temperature variations. As indicated in Sections 2.4.7 and 2.5, parts of the oceanic crust consist of serpentinite, that is, hydrated ultramafic rock. The ultramafic rock may be formed by magmatic differentiation within the gabbro layer, or derived directly from the mantle.

2.7 DIFFERENCES BETWEEN CONTINENTAL AND OCEANIC CRUST

On the basis of information presented in this and following chapters, the major differences between continental and oceanic crust can be summarized as follows:

1. *Layering*. The large-scale layering of the continental crust is ill defined and highly variable, reflecting a complex geologic history. In places there is a broad subdivision by the

Conrad discontinuity, but this is not globally developed. By contrast, the layering of the majority of oceanic crust is well defined into three distinct layers. However, the nature of these layers, in particular layers 2 and 3, may change quite markedly with depth.

2 *Thickness*. The thickness of continental crust averages 40 km but is quite variable, thinning to only a few kilometers beneath rifts and thickening to up to 80 km beneath young mountain belts. Most oceanic crust has a remarkably constant thickness of about 7 km, although layer 1, the sedimentary layer, increases in thickness towards ocean margins that are not characterized by ocean trenches. Differences in the thickness and the creep strength (Section 2.10.4) of continental crust make the lower crust of continental regions much more likely to deform pervasively than in the lower layers of oceanic crust (Section 2.10.5).

3 *Age*. Continental crust is as at least as old as 4.0 Ga, the age of the oldest rocks yet discovered (Section 11.1). On a very broad scale the oldest crust consists of Precambrian cratons or shield areas that are surrounded by younger orogenic belts, both active and inactive. Oceanic crust, however, is nowhere older than 180 Ma, and progressively increases in age outwards from oceanic ridges (Section 4.1). Oceans are consequently viewed as essentially transient features of the Earth's surface. About 50% of the surface area of the present day ocean floor has been created during the last 65 Ma, implying that 30% of the solid Earth's surface has been created during the most recent 1.5% of geologic time.

4 *Tectonic activity*. Continental crust may be extensively folded and faulted and preserves evidence of being subjected to multiple tectonic events. Oceanic crust, however, appears to be much more stable and has suffered relatively little deformation except at plate margins.

5 *Igneous activity*. There are very few active volcanoes on the great majority of the continental crust. The only major locations of activity are mountain belts of Andean type (Section 9.8). The activity within the oceans is very much greater. Ocean ridges and island arcs are the location of the Earth's most active areas of volcanic and plutonic activity. Oceanic islands are a third distinct, but less prolific oceanic setting for igneous activity.

2.8 THE MANTLE

2.8.1 Introduction

The mantle constitutes the largest internal subdivision of the Earth by both mass and volume, and extends from the Moho, at a mean depth of about 21 km, to the core–mantle boundary at a depth of 2891 km. On a gross scale it is believed to be chemically homogeneous, apart from the abundances of minor and trace elements, and formed of silicate minerals. The mineralogy and structure of the silicates change with depth and give rise to a transition zone between 410 and 660 km depth, which separates the upper and lower mantle.

Mantle materials are only rarely brought to the surface, in ophiolite complexes (Section 2.5), in kimberlite pipes (Section 13.2.2), and as xenoliths in alkali basalts. Consequently, most of our information about the mantle is indirect and based on the variation of seismic velocities with depth combined with studies of mineral behavior at high temperatures and pressure, and in shock-wave experiments. Geochemical studies of meteorites and ultramafic rocks are also utilized in making predictions about the mantle.

2.8.2 Seismic structure of the mantle

The uppermost part of the mantle constitutes a high velocity lid typically 80–160 km thick in which seismic velocities remain constant at a figure in excess of 7.9 km s^{-1} or increase slightly with depth. This part of the mantle makes up the lower portion of the lithosphere (Section 2.12). Beneath the lithosphere lies a *low velocity zone* extending to a depth of approximately 300 km. This appears to be present beneath most regions of the Earth with the exception of the mantle beneath cratonic areas. From the base of this zone seismic velocities increase slowly until a major discontinuity is reached at a depth of 410 km, marking the upper region

of the *transition zone*. There is a further velocity discontinuity at a depth of 660 km, the base of the transition zone.

Within the lower mantle velocities increase slowly with depth until the basal 200–300 km where gradients decrease and low velocities are present. This lowermost layer, at the core–mantle boundary, is known as *Layer D″* (Section 12.8.4) (Knittle & Jeanloz, 1991). Seismic studies have detected strong lateral heterogeneities and the presence of thin (5–50 km thick) *ultra-low velocity zones* at the base of Layer D″ (Garnero *et al.*, 1998).

2.8.3 Mantle composition

The fact that much of the oceanic crust is made up of material of a basaltic composition derived from the upper mantle suggests that the upper mantle is composed of either peridotite or eclogite (Harrison & Bonatti, 1981). The main difference between these two rock types is that peridotite contains abundant olivine and less than 15% garnet, whereas eclogite contains little or no olivine and at least 30% garnet. Both possess a seismic velocity that corresponds to the observed upper mantle value of about 8 km s^{-1}.

Several lines of evidence now suggest very strongly that the upper mantle is peridotitic. Beneath the ocean basins the P_n velocity is frequently anisotropic, with velocities over 15% higher perpendicular to ocean ridges. This can be explained by the preferred orientation of olivine crystals, whose long [100] axes are believed to lie in this direction. None of the common minerals of eclogite exhibit the necessary crystal elongation. A peridotitic composition is also indicated by estimates of Poisson's ratio from P and S velocities, and the presence of peridotites in the basal sections of ophiolite sequences and as nodules in alkali basalts. The density of eclogites is also too high to explain the Moho topography of isostatically compensated crustal structures.

The bulk composition of the mantle can be estimated in several ways: by using the compositions of various ultramafic rock types, from geochemical computations, from various meteorite mixtures, and by using data from experimental studies. It is necessary to distinguish between undepleted mantle and depleted mantle which has undergone partial melting so that many of the elements which do not easily substitute within mantle minerals have been removed and combined into the crust. The latter, so called "incompatible" elements, include the heat producing elements K, Th, and U. It is clear from the composition of mid-ocean ridge basalts (MORB), however, that the mantle from which they are derived by partial fusion is relatively depleted in these elements. So much so that, if the whole mantle had this composition, it would only account for a small fraction of the heat flow at the Earth's surface emanating from the mantle (Hofmann, 1997). This, and other lines of geochemical evidence, have led geochemists to conclude that all or most of the lower mantle must be more enriched in incompatible elements than the upper mantle and that it is typically not involved in producing melts that reach the surface. However, seismological evidence relating to the fate of subducted oceanic lithosphere (Sections 9.4, 12.8.2) and the lateral heterogeneity of Layer D″ suggests mantle wide convection and hence mixing (Section 12.9). Helffrich & Wood (2001) consider that the various lines of geochemical evidence can be reconciled with whole mantle convection if various small- and large-scale heterogeneities in the lower mantle revealed by seismological studies are remnants of subducted oceanic and continental crust. They estimate that these remnants make up about 16% and 0.3% respectively of the mantle volume.

Although estimates of bulk mantle composition vary in detail, it is generally agreed that at least 90% of the mantle by mass can be represented in terms of the oxides FeO, MgO, and SiO_2, and a further 5–10% is made up of CaO, Al_2O_3, and Na_2O.

2.8.4 The mantle low velocity zone

The low velocity zone (Fig. 2.16) is characterized by low seismic velocities, high seismic attenuation, and a high electrical conductivity. The seismic effects are more pronounced for S waves than for P waves. The low seismic velocities could arise from a number of different mechanisms, including an anomalously high temperature, a phase change, a compositional change, the presence of open cracks or fissures, and partial melting. All but the latter appear to be unlikely, and it is generally accepted that the lower seismic velocities arise because of the presence of molten material. That melting is likely to occur in this region is supported by the fact that it is at this level that mantle material

most closely approaches its melting point (Section 2.12, Fig. 2.36).

Only a very small amount of melt is required to lower the seismic velocity of the mantle to the observed values and to provide the observed attenuation properties. A liquid fraction of less than 1% would, if distributed along a network of fissures at grain boundaries, produce these effects (O'Connell & Budiansky, 1977). The melt may also be responsible for the high electrical conductivity of this zone. For the partial melting to occur, it is probable that a small quantity of water is required to lower the silicate melting point, and that this is supplied from the breakdown of hydrous mantle phases. The base of the low velocity zone and even its existence may be controlled by the availability of water in the upper mantle (Hirth & Kohlstedt, 2003).

The mantle low velocity zone is of major importance to plate tectonics as it represents a low viscosity layer along which relative movements of the lithosphere and asthenosphere can be accommodated.

2.8.5 The mantle transition zone

There are two major velocity discontinuities in the mantle at depths of 410 km and 660 km. The former marks the top of the transition zone and the latter its base. The discontinuities are rarely sharp and occur over a finite range in depth, so it is generally believed that they represent phase changes rather than changes in chemistry. Although these discontinuities could be due to changes in the chemical composition of the mantle at these depths, pressure induced phase changes are considered to be the more likely explanation. High-pressure studies have shown that olivine, the dominant mineral in mantle peridotite, undergoes transformations to the spinel structure at the pressure/temperature conditions at 410 km depth and then to perovskite plus magnesiowüstite at 660 km (Table 2.4) (Helffrich & Wood, 2001). Within subducting lithosphere, where the temperature at these depths is colder than in normal mantle, the depths at which these discontinuities occur are displaced exactly as predicted by thermal modeling and high-pressure experiments (Section 9.5). This lends excellent support to the hypothesis that the upper and lower bounds of the transition zone are defined by phase transformations. The other components of mantle peridotite, pyroxene and garnet, also undergo phase changes in this depth range but they are gradual and do not produce discontinuities in the variation of seismic velocity with depth. Pyroxene transforms into the garnet structure at pressures corresponding to 350–500 km depth; at about 580 km depth Ca-perovskite begins to exsolve from the garnet, and at 660–750 km the remaining garnet dissolves in the perovskite phase derived from the transformation of olivine. Thus the lower mantle mostly consists of phases with perovskite structure.

2.8.6 The lower mantle

The lower mantle represents approximately 70% of the mass of the solid Earth and almost 50% of the mass of the entire Earth (Schubert *et al.*, 2001). The generally smooth increase in seismic wave velocities with depth in most of this layer led to the assumption that it is relatively homogeneous in its mineralogy, having mostly a perovskite structure. However, more detailed seismo-

Table 2.4 *Phase transformations of olivine that are thought to define the upper mantle transition zone (after Helffrich & Wood, 2001).*

Depth	*Pressure*	
410 km	13–14 GPa	$(Mg,Fe)_2SiO_4 = (Mg,Fe)_2SiO_4$ Olivine Wadsleyite (β-spinel structure)
520 km	18 GPa	$(Mg,Fe)_2SiO_4 = (Mg,Fe)_2SiO_4$ Wadsleyite Ringwoodite (γ-spinel structure)
660 km	23 GPa	$(Mg,Fe)_2SiO_4 = (Mg,Fe)SiO_3 + (Mg,Fe)O$ Ringwoodite Perovskite Magnesiowüstite

logical studies have revealed that the lower mantle has thermal and/or compositional heterogeneity, probably as a result of the penetration of subducted oceanic lithosphere through the 660 km discontinuity (Section 2.8.3).

The lowest 200–300 km of the mantle, Layer D″ (Section 12.8.4), is often characterized by a decrease in seismic velocity, which is probably related to an increased temperature gradient above the mantle-core boundary. This lower layer shows large lateral changes in seismic velocity, indicating it is very heterogeneous. Ultra-low velocity zones, which show a 10% or greater reduction in both P and S wave velocities relative to the surrounding mantle, have been interpreted to reflect the presence of partially molten material (Williams & Garnero, 1996). These zones are laterally very heterogeneous and quite thin (5–40 km vertical thickness). Laboratory experiments suggest that the liquid iron of the core reacts with mantle silicates in Layer D″, with the production of metallic alloys and nonmetallic silicates from perovskite. Layer D″ thus is important because it governs core–mantle interactions and also may be the source of deep mantle plumes (Sections 12.8.4, 12.10).

2.9 THE CORE

The core, a spheroid with a mean radius of 3480 km, occurs at a depth of 2891 km and occupies the center of the Earth. The core–mantle boundary (Gutenberg discontinuity) generates strong seismic reflections and thus probably represents a compositional interface.

The outer core, at a depth of 2891–5150 km, does not transmit S waves and so must be fluid. This is confirmed by the generation of the geomagnetic field in this region by dynamic processes and by the long period variations observed in the geomagnetic field (Section 3.6.4). The convective motions responsible for the geomagnetic field involve velocities of $\sim 10^4$ m a^{-1}, five orders of magnitude greater than convection in the mantle. A fluid state is also indicated by the response of the Earth to the gravitational attraction of the Sun and Moon.

The boundary between the outer core and inner core at 5150 km depth is sharp, and not represented by any form of transition zone. The inner core is believed to be solid for several reasons. Certain oscillations of the Earth, produced by very large earthquakes, can only be explained by a solid inner core. A seismic phase has been recognized that travels to and from the inner core as a P wave, but traverses the inner core as an S wave. The amplitude of a phase reflected off the inner core also suggests that it must have a finite rigidity and thus be a solid.

Shock wave experiments have shown that the major constituents of both the inner and outer core must comprise elements of an atomic number greater than 23, such as iron, nickel, vanadium, or cobalt. Of these elements, only iron is present in sufficient abundance in the solar system to form the major part of the core. Again, by considering solar system abundances, it appears that the core should contain about 4% nickel. This iron–nickel mixture provides a composition for the outer core that is 8–15% too dense and it must therefore contain a small quantity of some lighter element or elements. The inner core, however, has a seismic velocity and density consistent with a composition of pure iron.

There are several candidates for the light elements present in the outer core, which include silicon, sulfur, oxygen, and potassium (Brett, 1976). Silicon requires an over-complex model for the formation of the Earth and sulfur conflicts with the idea that the interior of the Earth is highly depleted in volatile elements. Oxygen appears to be the most likely light element as FeO is probably sufficiently soluble in iron. The presence of potassium is speculative, but is interesting in that it would provide a heat source in the core that would be active over the whole of the Earth's history. It would also help to explain an apparent potassium deficiency in the Earth compared to meteorites.

2.10 RHEOLOGY OF THE CRUST AND MANTLE

2.10.1 Introduction

Rheology is the study of deformation and the flow of materials under the influence of an applied stress (Ranalli, 1995). Where temperature, pressure, and the magnitudes of the applied stresses are relatively low, rocks tend to break along discrete surfaces to form

fractures and faults. Where these factors are relatively high rocks tend to deform by ductile flow. Measures of strain are used to quantify the deformation.

Stress (σ) is defined as the force exerted per unit area of a surface, and is measured in Pascals (Pa). Any stress acting upon a surface can be expressed in terms of a normal stress perpendicular to the surface and two components of shear stress in the plane of the surface. The state of stress within a medium is conveniently specified by the magnitudes and directions of three *principal stresses* that act on three planes in the medium along which the shear stress is zero. The principal stresses are mutually orthogonal and are termed σ_1, σ_2, and σ_3, referring to the maximum, intermediate and minimum principal stresses, respectively. In the geosciences, compressive stresses are expressed as positive and tensile stresses negative. The magnitude of the difference between the maximum and minimum principal stresses is called the *differential stress*. *Deviatoric stress* represents the departure of a stress field from symmetry. The value of the differential stress and the characteristics of deviatoric stress both influence the extent and type of distortion experienced by a body.

Strain (ε) is defined as any change in the size or shape of a material. Strains are usually expressed as ratios that describe changes in the configuration of a solid, such as the change in the length of a line divided by its original length. *Elastic* materials follow Hooke's law where strain is proportional to stress and the strain is reversible until a critical stress, known as the *elastic limit*, is reached. This behavior typically occurs at low stress levels and high strain rates. Beyond the elastic limit, which is a function of temperature and pressure, rocks deform by either brittle fracturing or by ductile flow. The *yield stress* (or yield strength) is the value of the differential stress above the elastic limit at which deformation becomes permanent. *Plastic* materials display continuous, irreversible deformation without fracturing.

The length of time over which stress is applied also is important in the deformation of Earth materials (Park, 1983). Rock rheology in the short term (seconds or days) is different from that of the same material stressed over durations of months or years. This difference arises because rocks exhibit higher strength at high strain rates than at low strain rates. For example, when a block of pitch is struck with a hammer, that is, subjected to rapid "instantaneous" strain, it shatters. However, when left for a period of months, pitch deforms slowly by flowing. This slow long-term flow of materials under constant stress is known as *creep*. On time scales of thousands of years, information about the strength and rheology of the lithosphere mainly comes from observations of isostasy and lithospheric flexure (Section 2.11.4). On time scales of millions of years, Earth rheology generally is studied using a continuum mechanics approach, which describes the macroscopic relationships between stress and strain, and their time derivatives. Alternatively, the long-term rheology of the Earth may be studied using a microphysical approach, where the results of laboratory experiments and observations of microstructures are used to constrain the behavior of rocks. Both of these latter approaches have generated very useful results (e.g. Sections 7.6.6, 8.6.2, 10.2.5).

2.10.2 Brittle deformation

Brittle fracture is believed to be caused by progressive failure along a network of micro- and meso-scale cracks. The cracks weaken rock by producing local high concentrations of tensile stress near their tips. The crack orientations relative to the applied stress determine the location and magnitude of local stress maxima. Fracturing occurs where the local stress maxima exceed the strength of the rock.

This theory, known as the Griffith theory of fracture, works well under conditions of applied tensile stress or where one of the principal stresses is compressional. When the magnitude of the tensile stress exceeds the tensile strength of the material, cracks orthogonal to this stress fail first and an extension fracture occurs. Below a depth of a few hundred meters, where all principal stresses are usually compressional, the behavior of cracks is more complex. Cracks close under compression and are probably completely closed at depths of >5 km due to increasing overburden pressure. This implies that the compressive strength of a material is much greater than the tensile strength. For example, the compressive strength of granite at atmospheric pressure is 140 MPa, and its tensile strength only about 4 MPa.

Where all cracks are closed, fracturing depends upon the inherent strength of the material and the magnitude of the differential stress (Section 2.10.1). Experiments show that shear fractures, or faults, preferentially form at angles of $<45°$ on either side of the maximum principal compressive stress when a critical shear stress on the planes is exceeded. This critical shear stress (σ_s^*) depends upon the normal stress (σ_n) on planes of potential failure and the coefficient of internal friction (μ) on those planes, which resists relative motion across them.

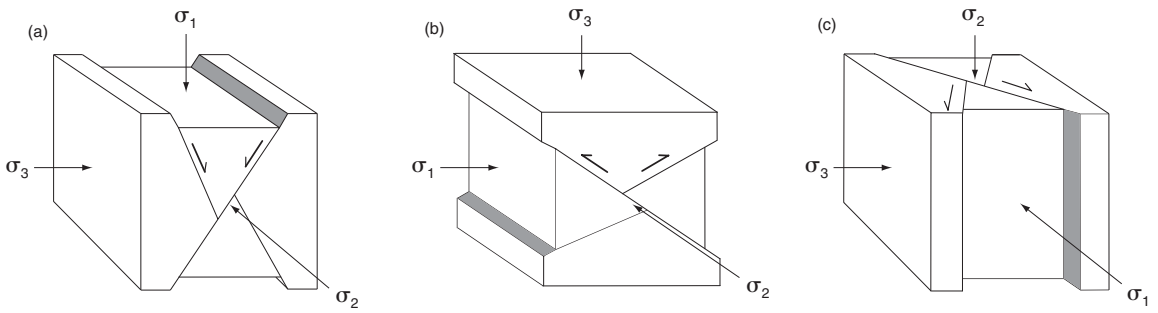


Figure 2.21 *Three classes of fault determined by the orientation of the principal stresses: (a) normal fault; (b) thrust fault; (c) strike-slip fault (after Angelier, 1994, with permission from Pergamon Press. Copyright Elsevier 1994).*

This relationship, called the Mohr–Coulomb fracture criterion, is described by the following linear equation:

$$|\sigma_s^*| = c + \mu\sigma_n$$

The cohesion (c) describes the resistance of the material to shear fracture on a plane of zero normal stress. Byerlee (1978) showed that many rock types have nearly the same coefficient of friction, within the range 0.6–0.8. The form of the equation, which is written using the absolute value of the critical shear stress, allows a pair of fractures to form that is symmetric about the axis of maximum principal compressive stress. Pore fluid pressure enhances fracturing by reducing the frictional coefficient and counteracting the normal stresses (σ_n) across the fault. The effect of pore fluid pressure explains faulting at depth, which would otherwise appear to require very high shear stresses because of the high normal stresses.

Under this compressional closed crack regime, the type of faulting which results, according to the theory of Anderson (1951), depends upon which of the principal stresses is vertical (Fig. 2.21). Normal, strike-slip, and thrust faults occur depending on whether σ_1, σ_2 or σ_3 respectively, is vertical. This theory is conceptually useful. However, it does not explain the occurrence of some faults, such as low-angle normal faults (Section 7.3), which display dips of ≤30°, flat thrust faults, or faults that develop in previously fractured, anisotropic rock.

The strength of rock increases with the pressure of the surrounding rock, termed the *confining pressure*, but decreases with temperature. In the uppermost 10–15 km of the crust the former effect is dominant and rock strength tends to increase with depth. Confining pressure increases with depth at a rate of about 33 MPa km^{-1} depending on the density of the overlying rocks. Below 10–15 km the effect of temperature takes over, and rocks may progressively weaken downwards. However, this simple relationship can be complicated by local variations in temperature, fluid content, rock composition, and pre-existing weaknesses.

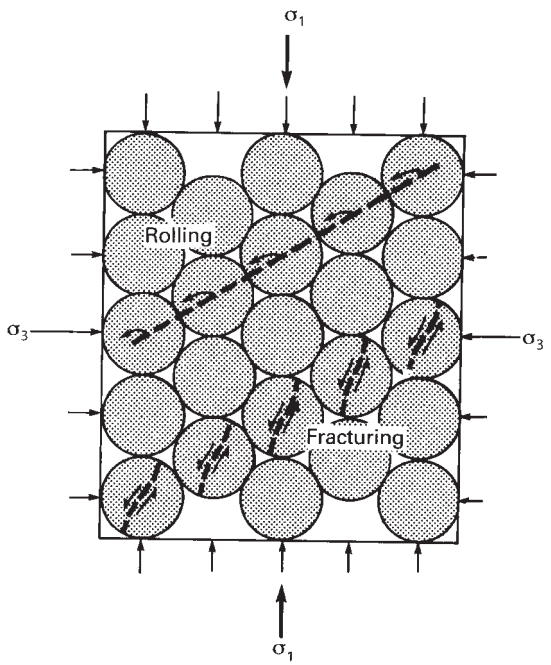


Figure 2.22 *Deformation of a brittle solid by cataclastic flow (redrawn from Ashby & Verrall, 1977, with permission from the Royal Society of London).*

The deformation of brittle solids can take the form of *cataclasis* (Fig. 2.22) (Ashby & Verrall, 1977). This

results from repeated shear fracturing, which acts to reduce the grain size of the rock, and by the sliding or rolling of grains over each other.

2.10.3 Ductile deformation

The mechanisms of ductile flow in crystalline solids have been deduced from studies of metals, which have the advantage that they flow easily at low temperatures and pressures. In general, where the temperature of a material is less than about half its melting temperature (T_m in Kelvin), materials react to low stresses by flowing slowly, or *creeping*, in the solid state. At high temperatures and pressures, the strength and flow of silicate minerals that characterize the crust (Tullis, 2002) and mantle (Li *et al.*, 2004) have been studied using experimental apparatus.

There are several types of ductile flow that may occur in the crust and mantle (Ashby & Verrall, 1977). All are dependent upon the ambient temperature and, less markedly, pressure. Increased temperature acts to lower the apparent viscosity and increase the strain rate, while increased pressure produces a more sluggish flow. In general, for ductile flow, the differential stress ($\Delta\sigma$) and the strain rate ($\delta\varepsilon/\delta\tau$) are related through a flow law of the form:

$$\Delta\sigma = [(\delta\varepsilon/\delta\tau)/A]^{1/n} \exp[E/nRT],$$

where E is the activation energy of the assumed creep process, T is temperature, R is the universal gas constant, n is an integer, and A is an experimentally determined constant.

Plastic flow occurs when the yield strength of the material is exceeded. Movement takes place by the gliding motions of large numbers of defects in the crystal lattices of minerals. Slip within a crystal lattice occurs as the individual bonds of neighboring atoms break and reform across glide planes (Fig. 2.23). This process results in linear defects, called *dislocations*, that separate slipped from unslipped parts of the crystal. The yield strength of materials deforming in this way is controlled by the magnitude of the stresses required to overcome the resistance of the crystal framework to the movement of the dislocations. The strain produced tends to be limited by the density of dislocations. The higher the density, the more difficult it is for dislocations to move in a process known as strain- or work-hardening.

Power-law creep (also known as *dislocation creep*) takes place at temperatures in excess of $0.55\,T_m$. In this form of creep the strain rate is proportional to the nth power of the stress, where $n \geq 3$. Power-law creep is similar to plastic flow, where deformation takes place by *dislocation glide*. However, in addition, the diffusion of atoms and of sites unoccupied by atoms called vacancies is permitted by the higher temperatures (Fig. 2.24). This diffusive process, termed *dislocation climb*, allows barriers to dislocation movement to be removed as they form. As a result work-hardening does not occur and steady state creep is facilitated. This balance results in *dynamic recrystallization* whereby new crystal grains form from old grains. Because of the higher temperature the yield strength is lower than for plastic flow, and strain results from lower stresses. Power-law creep is believed to be an important form of deformation in the upper mantle where it governs convective flow (Weertman, 1978). Newman & White (1997) suggest that the rheology of continental lithosphere is controlled by power-law creep with a stress exponent of three.

Diffusion creep dominates as temperatures exceed $0.85\,T_m$, and results from the migration of individual atoms and vacancies in a stress gradient (Fig. 2.25). Where the migration occurs through a crystal lattice it is known as *Nabarro–Herring creep*. Where it occurs along crystal boundaries it is known as *Coble creep*. In both forms of creep the strain rate ($\delta\varepsilon/\delta\tau$) is proportional to the differential stress ($\Delta\sigma$) with the constant of proportionality being the dynamic viscosity (η). This relationship is given by:

$$\Delta\sigma = 2\eta(\delta\varepsilon/\delta\tau)$$

The viscosity increases as the square of the grain radius so that a reduction in grain size is expected to result in rheological weakening. Diffusion creep is believed to occur in the asthenosphere (Section 2.12) and in the lower mantle (Section 2.10.6).

Superplastic creep has been observed in metals and may also occur in some rocks. This type of creep results from the coherent sliding of crystals along grain boundaries where the movement occurs without opening up gaps between grains. The sliding may be accommodated by both diffusion and dislocation mechanisms. Superplastic creep is characterized by a power-law rheology with a stress exponent of one or two and is associated with high strain rates. Some studies (e.g. Karato, 1998) have inferred that superplastic creep contributes

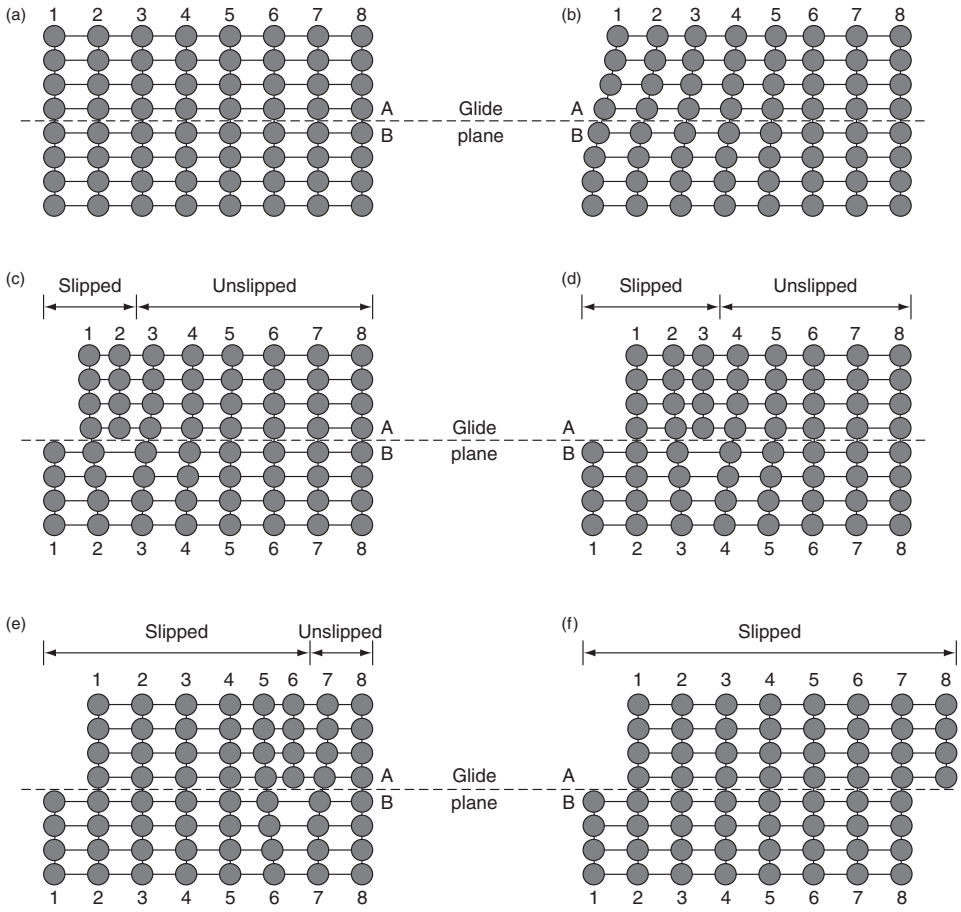


Figure 2.23 *Plastic flow by the migration of a linear edge dislocation through a crystal (from Structural Geology by Robert J. Twiss and Eldridge M. Moores. © 1992 by W.H. Freeman and Company. Used with permission).*

to deformation in the lower mantle, although this interpretation is controversial.

2.10.4 Lithospheric strength profiles

In most quantitative treatments of deformation at large scales, the lithosphere is assumed to consist of multiple layers characterized by different rheologies (e.g. Section 7.6.6). The rheologic behavior of each layer depends on the level of the differential stress ($\Delta\sigma$) and the lesser of the calculated brittle and ductile yield stresses (Section 2.10.1). The overall strength of the lithosphere and its constituent layers can be estimated by integrating yield stress with respect to depth. This integrated strength is highly sensitive to the geothermal gradient as well as to the composition and thickness of each layer, and to the presence or absence of fluids.

The results of deformation experiments and evidence of compositional variations with depth (Section 2.4) have led investigators to propose that the lithosphere is characterized by a "jelly sandwich" type rheological layering (Ranalli & Murphy, 1987), where strong layers separate one or more weak layers. For example, Brace & Kohlstedt (1980) investigated the limits of lithospheric strength based on measurements on quartz and olivine, which are primary constituents of the

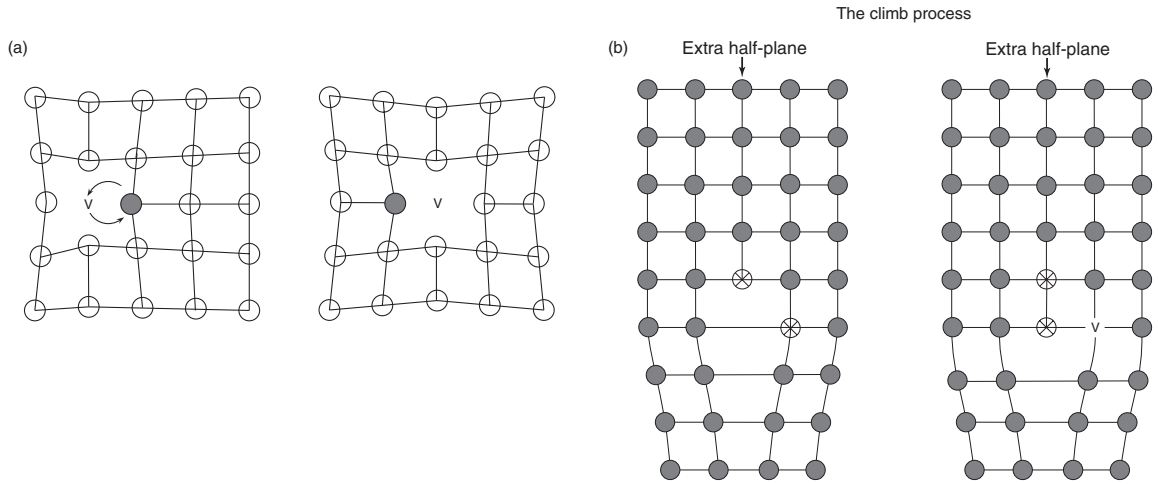


Figure 2.24 *(a) The diffusion of a vacancy (v) through a crystal; (b) the downward climb of an edge dislocation as adjacent atoms (crossed) exchange bonds leaving behind a vacancy that moves by diffusion (from Structural Geology by Robert J. Twiss and Eldridge M. Moores. © 1992 by W.H. Freeman and Company. Used with permission).*

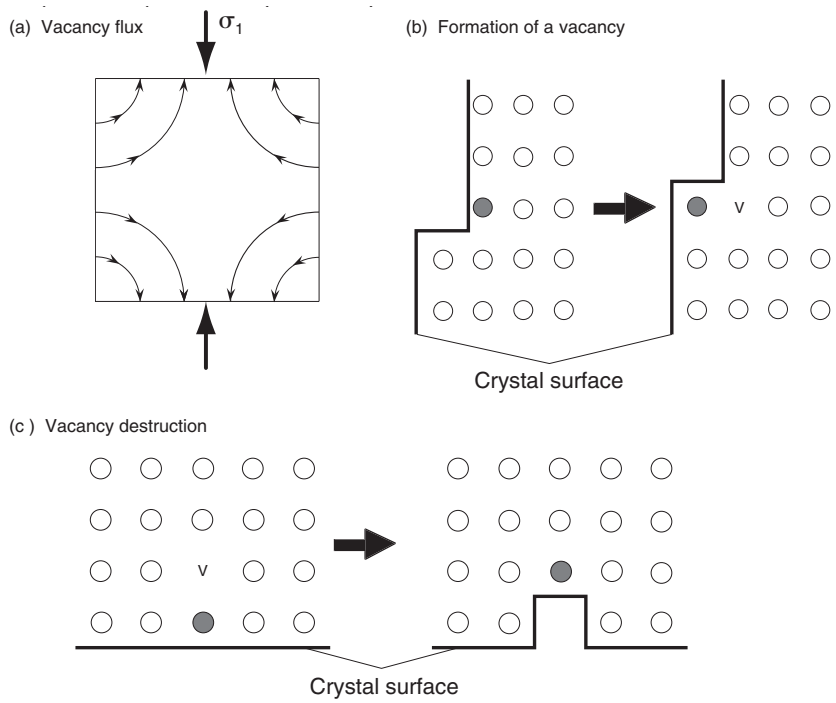


Figure 2.25 *Nabarro–Herring creep: (a) vacancies diffuse toward surfaces of high normal stress; (b) creation of a vacancy (v) at a surface of minimum compressive stress; (c) destruction of a vacancy at a surface of maximum compressive stress (from Structural Geology by Robert J. Twiss and Eldridge M. Moores. © 1992 by W.H. Freeman and Company. Used with permission). Solid lines in b and c mark crystal surface, solid circle marks the ion whose position changes during the creation of a vacancy.*

continental crust and upper mantle, respectively. The results of these and other measurements (e.g. Ranalli & Murphy, 1987; Mackwell *et al.*, 1998) suggest that within the oceanic lithosphere the upper brittle crust gives way to a region of high strength at a depth of 20–60 km, depending on the temperature gradient (Fig. 2.26a). Below this depth the strength gradually decreases and grades into that of the asthenosphere. Continental crust, however, is much thicker than oceanic crust, and at the temperatures of 400–700°C experienced in its lower layers the minerals are much weaker than the olivine found at these depths in the oceanic lithosphere. Whereas the oceanic lithosphere behaves as a single rigid plate because of its high strength, the continental lithosphere does not (Sections 2.10.5, 8.5) and typically is characterized by one or more layers of weakness at deep levels (Fig. 2.26b,c).

Figure 2.26c,d shows two other experimentally determined strength curves for continental lithosphere that illustrate the potential effects of water on the strength of various layers. These curves were calculated using rheologies for diabase and other crustal and mantle rocks, a strain rate of $(\delta\varepsilon/\delta\tau) = 10^{-15}\,s^{-1}$, a typical thermal gradient for continental crust with a surface heat flow of $60\,mW\,m^{-2}$, and a crustal thickness of 40 km (Mackwell *et al.*, 1998). The upper crust (0–15 km depth) is represented by wet quartz and Byerlee's (1978) frictional strength law (Section 2.10.2), and the middle crust (15–30 km depth) by wet quartz and power-law creep (Section 2.10.3). These and other postulated strength profiles commonly are used in thermomechanical models of continental deformation (Sections 7.6.6, 8.6.2, 10.2.5). However, it is important to keep in mind that the use of any one profile in a particular setting involves considerable uncertainty and is the subject of much debate (Jackson, J., 2002; Afonso & Ranalli, 2004; Handy & Brun, 2004). In settings where ambient conditions appear to change frequently, such as within orogens and magmatic arcs, several curves may be necessary to describe variations in rock strength with depth for different time periods.

2.10.5 Measuring continental deformation

Zones of continental deformation commonly are wider and more diffuse than zones of deformation affecting oceanic lithosphere. This characteristic results from the thickness, composition, and pressure–temperature profile of continental crust, which makes ductile flow in its lower parts more likely than it is in oceanic regions. The width and diffusivity of these zones make some of the concepts of plate tectonics, such as the rigid motion of plates along narrow boundaries, difficult to apply to the continents. Consequently, the analysis of continental deformation commonly requires a framework that is different to that used to study deformation in oceanic lithosphere (e.g. Section 8.5).

At the scale of large tectonic features such as wide intracontinental rifts (Section 7.3), continental transforms (Section 8.5), and orogenic belts (Section 10.4.3), deformation may be described by a regional horizontal velocity field rather than by the relative motion of rigid blocks (e.g. Fig. 8.18b). Methods of estimating the regional velocity field of deforming regions usually involves combining information from Global Positioning System (GPS) satellite measurements (Clarke *et al.*, 1998), fault slip rates (England & Molnar, 1997), and seismicity (Jackson *et al.*, 1992). One of the challenges of this approach is the short, decade-scale time intervals over which GPS data are collected. These short intervals typically include relatively few major earthquakes. Consequently, the measured surface motions mostly reflect nonpermanent, elastic strains that accumulate between major seismic events (i.e. interseismic) rather than the permanent strains that occur during ruptures (Bos & Spakman, 2005; Meade & Hager, 2005). This characteristic results in a regional velocity field that rarely shows the discontinuities associated with slip on major faults. Instead the displacements on faults are described as continuous functions and the velocity field is taken to represent the average deformation over a given region (Jackson, 2004). Nevertheless, regional velocity fields have proven to be a remarkably useful way of describing continental deformation. The methods commonly used to process and interpret them are discussed further in Sections 5.3 and 8.5.

Synthetic Aperture Radar (SAR) also is used to measure ground displacements, including those associated with volcanic and earthquake activity (Massonnet & Feigl, 1998). The technique involves using SAR data to measure small changes in surface elevations from satellites that fly over the same area at least twice, called repeat-pass Interferometric SAR, or InSAR. GPS data and strain meters provide more accurate and frequent observations of deformation in specific areas, but InSAR is especially good at revealing the spatial complexity of displacements that occur in tectonically active areas. In

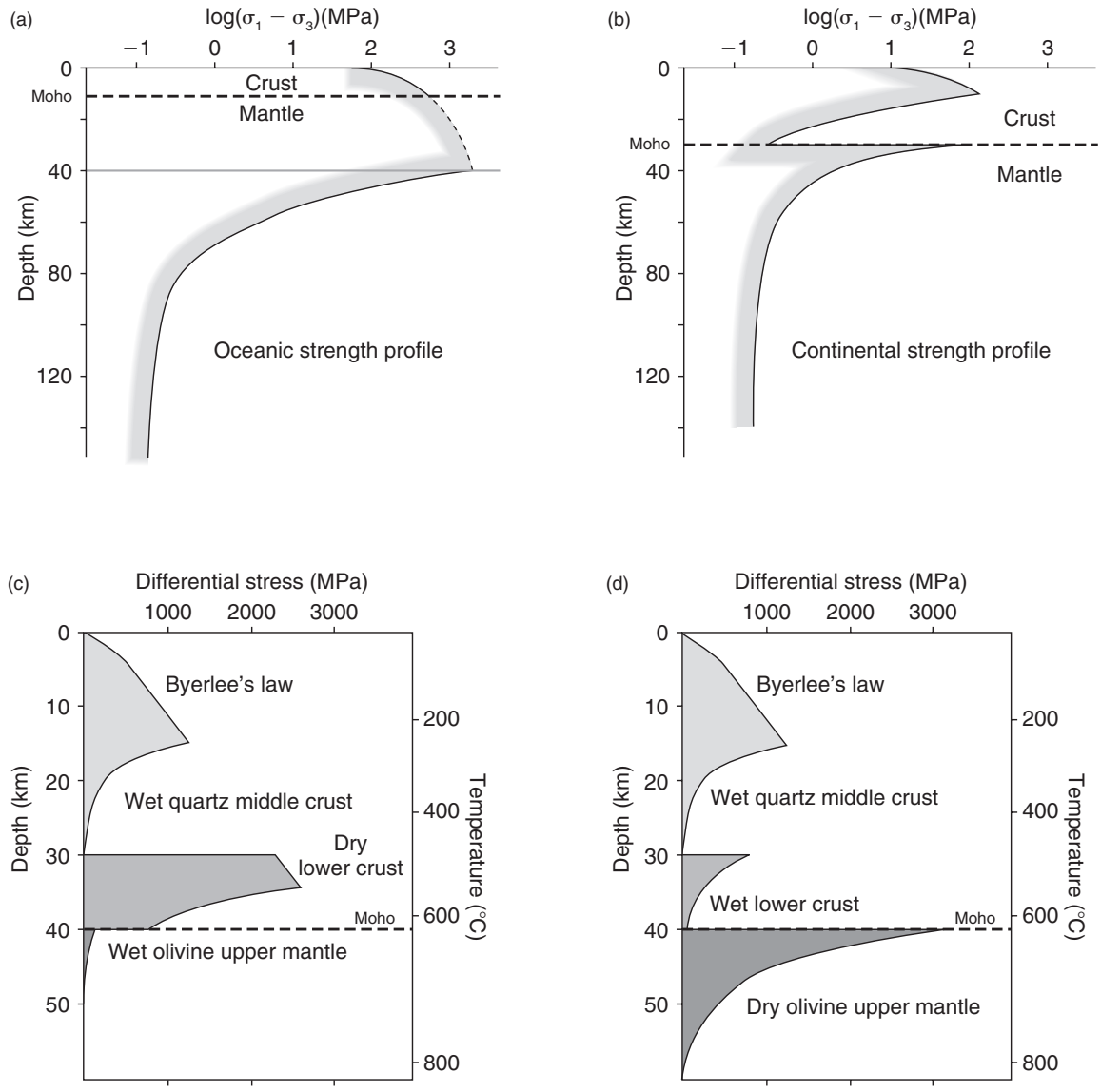


Figure 2.26 *Schematic strength profiles through (a) oceanic and (b) continental lithosphere (after Ranalli, 1995, fig. 12.2. Copyright © 1995, with kind permission of Springer Science and Business Media). Profile in (a) shows a 10-km-thick mafic crust and a 75-km-thick lithosphere. Profile in (b) shows a 30-km-thick unlayered crust and a thin, 50-km-thick lithosphere. Profiles in (c) and (d) incorporate a wet middle crust and show a dry lower crust and a wet upper mantle, and a wet lower crust and dry upper mantle, respectively (modified from Mackwell et al., 1998, by permission of the American Geophysical Union. Copyright © 1998 American Geophysical Union) See text for explanation.*

the central Andes and Kamchatka InSAR measurements have been used to evaluate volcanic hazards and the movement of magma in volcanic arcs (Pritchard & Simons, 2004). In southeast Iran, InSAR data have been used to determine the deformation field and source parameters of a magnitude $M_w = 6.5$ earthquake that affected the city of Bam in 2003 (Wang *et al.*, 2004). The combined use of GPS and InSAR data have revealed the vertical displacements associated with a part of the San Andreas Fault system near San Francisco (Fig. 8.7b).

2.10.6 Deformation in the mantle

Measurements of seismic anisotropy (Section 2.1.8) and the results of mineral physics experiments have been used to infer creep mechanisms and flow patterns in the mantle (Karato, 1998; Park & Levin, 2002; Bystricky, 2003). The deformation of mantle minerals, including olivine, by dislocation creep results in either a preferred orientation of crystal lattices or a preferred orientation of mineral shapes. This alignment affects how fast seismic waves propagate in different directions. Measurements of this directionality and other properties potentially allow investigators to image areas of the mantle that are deforming by dislocation creep (Section 2.10.3) and to determine whether the flow is mostly vertical or mostly horizontal. However, these interpretations are complicated by factors such as temperature, grain size, the presence of water and partial melt, and the amount of strain (Hirth & Kohlstedt, 2003; Faul *et al.*, 2004).

Most authors view power-law (or dislocation) creep as the dominant deformation mechanism in the upper mantle. Experiments on olivine, structural evidence in mantle-derived nodules, and the presence of seismic anisotropy suggest that power-law creep occurs to a depth of at least 200 km. These results contrast with many studies of post-glacial isostatic rebound (Section 2.11.5), which tend to favor a diffusion creep mechanism for flow in the upper mantle. Karato & Wu (1993) resolved this apparent discrepancy by suggesting that a transition from power-law creep to diffusion creep occurs with depth in the upper mantle. Diffusion creep may become increasingly prominent with depth as pressure and temperature increase and stress differences decrease. A source of potential uncertainty in studies of mantle rheology using glacial rebound is the role of *transient creep*, where the strain rate varies with time under constant stress. Because the total strains associated with rebound are quite small ($\leq 10^{-3}$) compared to the large strains associated with mantle convection, transient creep may be important during post-glacial isostatic rebound (Ranalli, 2001).

In contrast to the upper mantle, much of the lower mantle is seismically isotropic, suggesting that diffusion creep is the dominant mechanism associated with mantle flow at great depths (Karato *et al.*, 1995). Unlike dislocation creep, diffusion creep (and also superplastic creep) result in an isotropic crystal structure in lower mantle minerals, such as perovskite and magnesiowüstite. Large uncertainties about lower mantle rheology exist because lower mantle materials are difficult to reproduce in the laboratory. Nevertheless, advances in high-pressure experimentation have allowed investigators to measure some of the physical properties of lower mantle minerals. Some measurements suggest that lower mantle rheology strongly depends on the occurrence and geometry of minor, very weak phases, such as magnesium oxide (Yamazaki & Karato, 2001). Murakami *et al.* (2004) demonstrated that at pressure and temperature conditions corresponding to those near the core–mantle boundary, $MgSiO_3$ perovskite transforms to a high-pressure form that may influence the seismic characteristics of the mantle below the D″ discontinuity (Section 12.8.4).

Unlike most of the lower mantle, observations at the base of the mesosphere, in the D″ layer (Section 2.8.5), indicate the presence of seismic anisotropy (Panning & Romanowicz, 2004). The dominance of V_{SH} polarization over V_{SV} in shear waves implies large-scale horizontal flow, possibly analogous to that found in the upper 200 km of the mantle. The origin of the anisotropy, whether it is due to the alignment of crystal lattices or to the preferred orientation of mineral shapes, is uncertain. However, these observations suggest that D″ is a mechanical boundary layer for mantle convection. Exceptions to the pattern of horizontal flow at the base of the lower mantle are equally interesting. Two exceptions occur at the bottom of extensive low velocity regions in the lower mantle beneath the central Pacific and southern Africa (Section 12.8.2) where anisotropy measurements indicate the onset of vertical upwelling (Panning & Romanowicz, 2004).

Another zone of seismic anisotropy and horizontal flow similar to that in the D″ layer also may occur at

the top of the lower mantle or mesosphere (Karato, 1998). However, this latter interpretation is highly controversial and awaits testing by continued investigation. If such a zone of horizontal flow does exist then convection in the mantle probably occurs in layers and does not involve the whole mantle (Section 12.5.3).

2.11 ISOSTASY

2.11.1 Introduction

The phenomenon of isostasy concerns the response of the outer shell of the Earth to the imposition and removal of large loads. This layer, although relatively strong, is unable to support the large stresses generated by, for example, the positive weight of a mountain range or the relative lack of weight of an ocean basin. For such features to exist on the Earth's surface, some form of compensating mechanism is required to avoid the large stresses that would otherwise be generated.

Isostasy was first recognized in the 18th century when a party of French geodesists were measuring the length of a degree of latitude in Ecuador in an attempt to determine if the shape of the Earth corresponds to an oblate or a prolate ellipsoid. Plumb lines were used as a vertical reference in the surveying and it was recognized that a correction would have to be applied for the horizontal deflection caused by the gravitational attraction of the Andes. When this correction, based on the mass of the Andes above sea level, was applied, however, it was found that the actual vertical deflection was less than predicted (Fig. 2.27). This phenomenon was attributed to the existence of a negative mass anomaly beneath the Andes that compensates, that is to say, supports, the positive mass of the mountains. In the 19th century similar observations were made in the vicinity of the Himalaya and it was recognized that the compensation of surface loading at depth is a widespread phenomenon.

The presence of subsurface compensation is confirmed by the variation in the Earth's gravitational field over broad regions. Bouguer anomalies (Kearey *et al.*, 2002) are generally negative over elevated continental areas and positive over ocean basins (Fig. 2.28). These observations confirm that the positive topography of continents and negative topography of oceans is compensated by regions at depth with density contrasts which are, respectively, negative and positive and whose mass anomaly approximates that of the surface features.

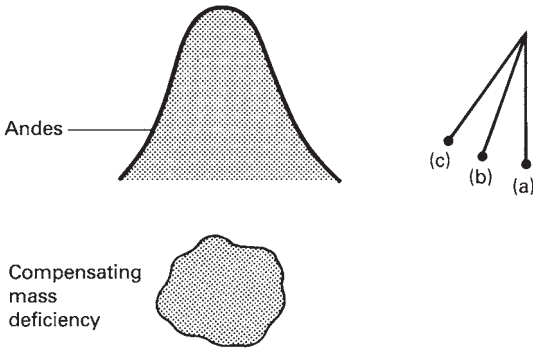


Figure 2.27 *Horizontal gravitational attraction of the mass of the Andes above sea level would cause the deflection (c) of a plumb bob from the vertical (a). The observed deflection (b) is smaller, indicating the presence of a compensating mass deficiency beneath the Andes (angles of deflection and mass distribution are schematic only).*

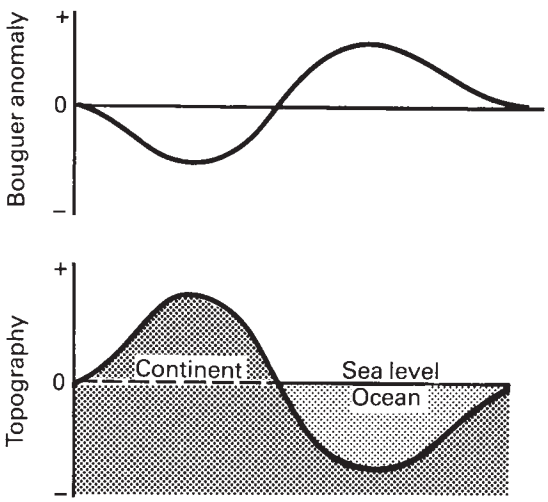


Figure 2.28 *Inverse correlation of Bouguer anomalies with topography indicating its isostatic compensation.*

The principle of isostasy is that beneath a certain depth, known as the depth of compensation, the pressures generated by all overlying materials are every-

where equal; that is, the weights of vertical columns of unit cross-section, although internally variable, are identical at the depth of compensation if the region is in isostatic equilibrium.

Two hypotheses regarding the geometric form of local isostatic compensation were proposed in 1855 by Airy and Pratt.

2.11.2 Airy's hypothesis

Airy's hypothesis assumes that the outermost shell of the Earth is of a constant density and overlies a higher density layer. Surface topography is compensated by varying the thickness of the outer shell in such a way that its buoyancy balances the surface load. A simple analogy would be blocks of ice of varying thickness floating in water, with the thickest showing the greatest elevation above the surface. Thus mountain ranges would be underlain by a thick root, and ocean basins by a thinned outer layer or antiroot (Fig. 2.29a). The base of the outer shell is consequently an exaggerated mirror image of the surface topography. Consider the columns of unit cross-section beneath a mountain range and a region of zero elevation shown in Fig. 2.29a. Equating their weights gives:

$$g[h\rho_c + T_A\rho_c + r\rho_c + D_A\rho_m] = g[T_A\rho_c + r\rho_m + D_A\rho_m]$$

where g is the acceleration due to gravity.

Rearranging this equation gives the condition for isostatic equilibrium:

$$r = \frac{h\rho_c}{(\rho_m - \rho_c)}$$

A similar computation provides the condition for compensation of an ocean basin:

$$a = \frac{z(\rho_c - \rho_w)}{(\rho_m - \rho_c)}$$

If one substitutes appropriate densities for the crust, mantle, and sea water in these equations they predict that the relief on the Moho should be approximately seven times the relief at the Earth's surface.

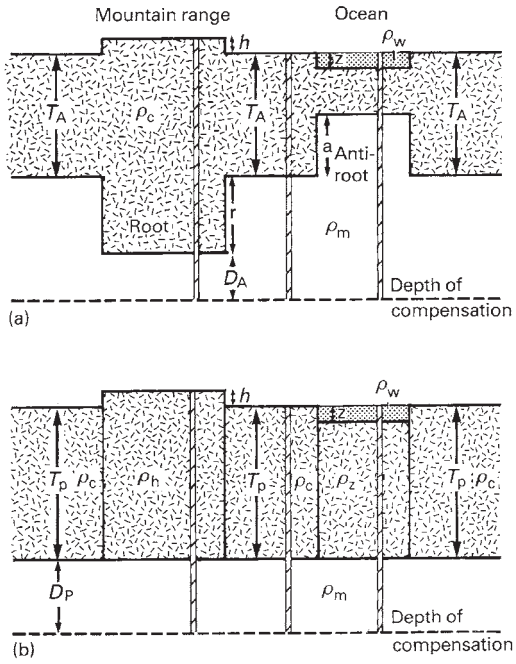


Figure 2.29 *(a) Airy mechanism of isostatic compensation. h, height of mountain above sea level; z, depth of water of density ρ_w; T_A, normal thickness of crust of density ρ_c; r, thickness of root; a, thickness of antiroot; D_A, depth of compensation below root; ρ_m, density of mantle. (b) The Pratt mechanism of isostatic compensation. Legend as for (a) except T_p, normal thickness of crust; ρ_h, density of crust beneath mountain; ρ_z, density of crust beneath ocean; D_p, depth of compensation below T_p.*

2.11.3 Pratt's hypothesis

Pratt's hypothesis assumes a constant depth to the base of the outermost shell of the Earth, whose density varies according to the surface topography. Thus, mountain ranges would be underlain by relatively low density material and ocean basins by relatively high density material (Fig. 2.29b). Equating the weights of columns of unit cross-section beneath a mountain range and a region of zero elevation gives:

$$g(T_p + h)\rho_h = gT_p\rho_c$$

which on rearrangement provides the condition for isostatic equilibrium of the mountain range:

$$\rho_h = \frac{T_p \rho_c}{(T_p + h)}$$

A similar computation for an ocean basin gives:

$$\rho_z = \frac{(T_p \rho_c - z\rho_w)}{(T_p - z)}$$

In these early models of isostasy it was assumed that the outer shell of the Earth, whose topography is compensated, corresponded to the crust. Certainly the large density contrast existing across the Moho plays a major part in the compensation. It is now believed, however, that the compensated layer is rather thicker and includes part of the upper mantle. This strong outer layer of the Earth is known as the lithosphere (Section 2.12). The lithosphere is underlain by a much weaker layer known as the asthenosphere which deforms by flow, and which can thus be displaced by vertical movements of the lithosphere. The density contrast across the lithosphere-asthenosphere boundary is, however, very small.

Both the Airy and Pratt hypotheses are essentially applications of Archimedes' Principle whereby adjacent blocks attain isostatic equilibrium through their buoyancy in the fluid substratum. They assume that adjacent blocks are decoupled by fault planes and achieve equilibrium by rising or subsiding independently. However, these models of *local* compensation imply unreasonable mechanical properties for the crust and upper mantle (Banks *et al.*, 1977), because they predict that independent movement would take place even for very small loads. The lithosphere is demonstrably not as weak as this implies, as large gravity anomalies exist over igneous intrusions with ages in excess of 100 Ma. The lithosphere must therefore be able to support stress differences of up to 20–30 MPa for considerable periods of time without the necessity of local compensation.

2.11.4 Flexure of the lithosphere

More realistic models of isostasy involve *regional* compensation. A common approach is to make the analogy between the lithosphere and the behavior of an elastic sheet under load. Figure 2.30 illustrates the elastic response to loading; the region beneath the load subsides over a relatively wide area by displacing asthenospheric material, and is complemented by the development of peripheral bulges. Over long periods of time, however, the lithosphere may act in a viscoelastic manner and undergo some permanent deformation by creep (Section 2.10.3).

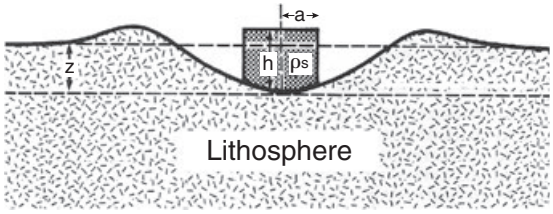


Figure 2.30 *Flexural downbending of the lithosphere as a result of a two-dimensional load of half-width a, height h, and density ρ_s.*

For example, the vertical displacement z of the oceanic lithosphere under loading can be calculated by modeling it as an elastic sheet by solving the fourth order differential equation:

$$D\frac{d^4 z}{dx^4} + (\rho_m - \rho_w) z g = P(x)$$

where $P(x)$ is the load as a function of horizontal distance x, g the acceleration due to gravity, and ρ_m, ρ_w the densities of asthenosphere and sea water, respectively. D is a parameter termed the flexural rigidity, which is defined by:

$$D = ET_e^3/12(1 - \sigma^2)$$

where E is Young's modulus, σ Poisson's ratio, and T_e the thickness of the elastic layer of the lithosphere.

The specific relationship between the displacement z and load for the two-dimensional load of half-width a, height h, and density ρ_s shown in Fig. 2.30 is:

$$z_{max} = h(\rho_s - \rho_w)(1 - e^{-\lambda a} \cos \lambda a)/(\rho_m - \rho_s)$$

where

$$\lambda = \sqrt[4]{(\rho_m - \rho_w) g / 4D}$$

and ρ_w, ρ_m the densities of water and the mantle, respectively.

Note that as the elastic layer becomes more rigid, D approaches infinity, λ approaches zero, and the depression due to loading becomes small. Conversely, as the layer becomes weaker, D approaches zero, λ approaches infinity, and the depression approaches $h(\rho_s - \rho_w)/(\rho_m - \rho_s)$ (Watts & Ryan, 1976). This is equivalent to Airy-type isostatic equilibrium and indicates that for this mechanism to operate the elastic layer and fluid substrate must both be very weak.

It can be shown that, for oceanic lithosphere away from mid-ocean ridges, loads with a half-width of less than about 50 km are supported by the finite strength of the lithosphere. Loads with half-widths in excess of about 500 km are in approximate isostatic equilibrium. Figure 2.31 illustrates the equilibrium attained by the oceanic lithosphere when loaded by a seamount (Watts *et al.*, 1975). Thus, as a result of its flexural rigidity, the lithosphere has sufficient internal strength to support relatively small loads without subsurface compensation. Such loads include small topographic features and variations in crustal density due, for example, to small granitic or mafic bodies within the crust. This more realistic model of isostatic compensation, that takes into account the flexural rigidity of the lithosphere, is referred to as *flexural isostasy* (Watts, 2001).

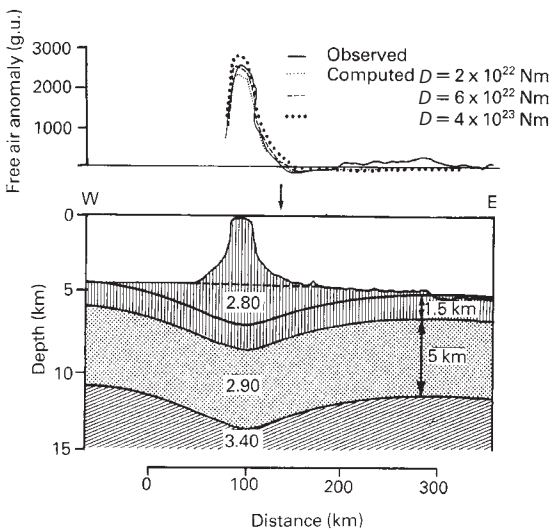


Figure 2.31 *Interpretation of the free air anomaly of the Great Meteor Seamount, northeast Atlantic Ocean, in terms of flexural downbending of the crust. A model with the flexural rigidity (D) of 6×10^{22} N m appears best to simulate the observed anomaly. Densities in Mg m^{-3}. Arrow marks the position 30°N, 28°W (redrawn from Watts et al., 1975, by permission of the American Geophysical Union. Copyright © 1975 American Geophysical Union).*

2.11.5 Isostatic rebound

The equilibrium flexural response of the lithosphere to loading is independent of the precise mechanical properties of the underlying asthenosphere as long as it facilitates flow. However, the reattainment of equilibrium after removal of the load, a phenomenon known as *isostatic rebound*, is controlled by the viscosity of the asthenosphere. Measurement of the rates of isostatic rebound provides a means of estimating the viscosity of the upper mantle. Fennoscandia represents an example of this type of study as precise leveling surveys undertaken since the late 19th century have shown that this region is undergoing uplift following the melting of the Pleistocene ice sheet (Fig. 2.32). The maximum uplift rates occur around the Gulf of Bothnia, where the land is rising at a rate of over 10 mm a^{-1}. Twenty thousand years ago the land surface was covered by an ice sheet about 2.5 km thick (Fig. 2.32a). The lithosphere accommodated this load by flexing (Fig. 2.32b), resulting in a subsidence of 600–700 m and a lateral displacement of asthenospheric material. This stage currently pertains in Greenland and Antarctica where, in Greenland, the land surface is depressed by as much as 250 m below sea level by the weight of ice. Melting of the ice was complete about 10,000 years ago (Fig. 2.32c), and since this time the lithosphere has been returning to its original position and the land rising in order to regain isostatic equilibrium. A similar situation pertains in northern Canada where the land surface around Hudson Bay is rising subsequent to the removal of an icecap. The rate of isostatic rebound provides an estimate for the viscosity of the upper mantle of 10^{21} Pa s (Pascal seconds), and measurements based on world-wide modeling of post-glacial recovery and its associated oceanic loading suggest that this figure generally applies throughout the upper mantle as a whole (Peltier & Andrews, 1976). Compared to the viscosity of water (10^{-3} Pa s) or a lava flow (4×10^{3} Pa s), the viscosity of the sub-lithospheric mantle is extremely high and its fluid behavior is only apparent in processes with a large

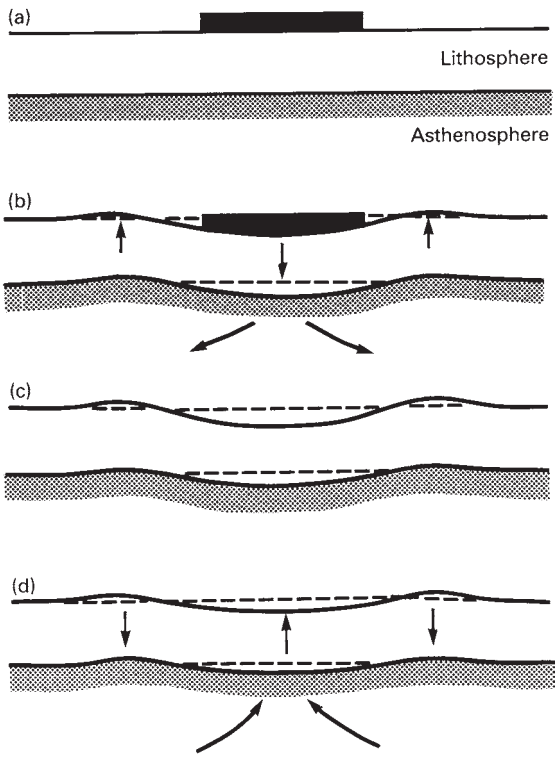


Figure 2.32 *Theory of isostatic rebound. (a) The load of an icecap on the lithosphere causes downbending accompanied by the elevation of the peripheral lithosphere and lateral flow in the asthenosphere (b). When the icecap melts (c), isostatic equilibrium is regained by reversed flow in the asthenosphere, sinking of the peripheral bulges and elevation of the central region (d).*

time constant. Knowledge of the viscosity of the mantle, however, provides an important control on the nature of mantle convection, as will be discussed in Section 12.5.2.

2.11.6 Tests of isostasy

The state of isostatic compensation of a region can be assessed by making use of gravity anomalies. The *isostatic anomaly*, IA, is defined as the Bouguer anomaly minus the gravity anomaly of the subsurface compensation. Consider a broad, flat plateau of elevation h compensated by a root of thickness r. The terrain correction of such a feature is small in the central part of the plateau so that here the Bouguer anomaly, BA, is related to the free-air anomaly, FAA by the relationship:

$$BA = FAA - BC$$

where BC is the Bouguer correction, equal to $2\pi G\rho_c h$, where ρ_c is the density of the compensated layer. For such an Airy compensation:

$$IA = BA - A_{root}$$

where A_{root} is the gravity anomaly of the compensating root. Since the root is broad compared to its thickness, its anomaly may be approximated by that of an infinite slab, that is $2\pi G(\rho_c - \rho_m)r$, where ρ_m is the density of the substrate. Combining the above two equations:

$$IA = FAA - 2\pi G\rho_c h - 2\pi G(\rho_c - \rho_m)r$$

From the Airy criterion for isostatic equilibrium:

$$r = h\rho_c/(\rho_m - \rho_c)$$

Substitution of this condition into the equation reveals that the isostatic anomaly is equal to the free-air anomaly over a broad flat feature, and this represents a simple method for assessing the state of isostatic equilibrium. Figure 2.33 shows free-air, Bouguer and isostatic anomalies over a broad flat feature with varying degrees of compensation. Although instructive in illustrating the similarity of free-air and isostatic anomalies, and the very different nature of the Bouguer anomaly, this simple Airy isostatic anomaly calculation is clearly unsatisfactory in not taking into account topography and regional compensation due to flexure of the lithosphere.

To test isostasy over topographic features of irregular form more accurate computation of isostatic anomalies is required. This procedure involves calculating the shape of the compensation required by a given hypothesis of isostasy, computing its gravity anomaly, and then subtracting this from the observed Bouguer anomaly to provide the isostatic anomaly. The technique of computing the gravity anomaly from a hypothetical model is known as *forward modeling*.

Gravity anomalies can thus be used to determine if a surface feature is isostatically compensated at depth. They cannot, however, reveal the form of compensa-

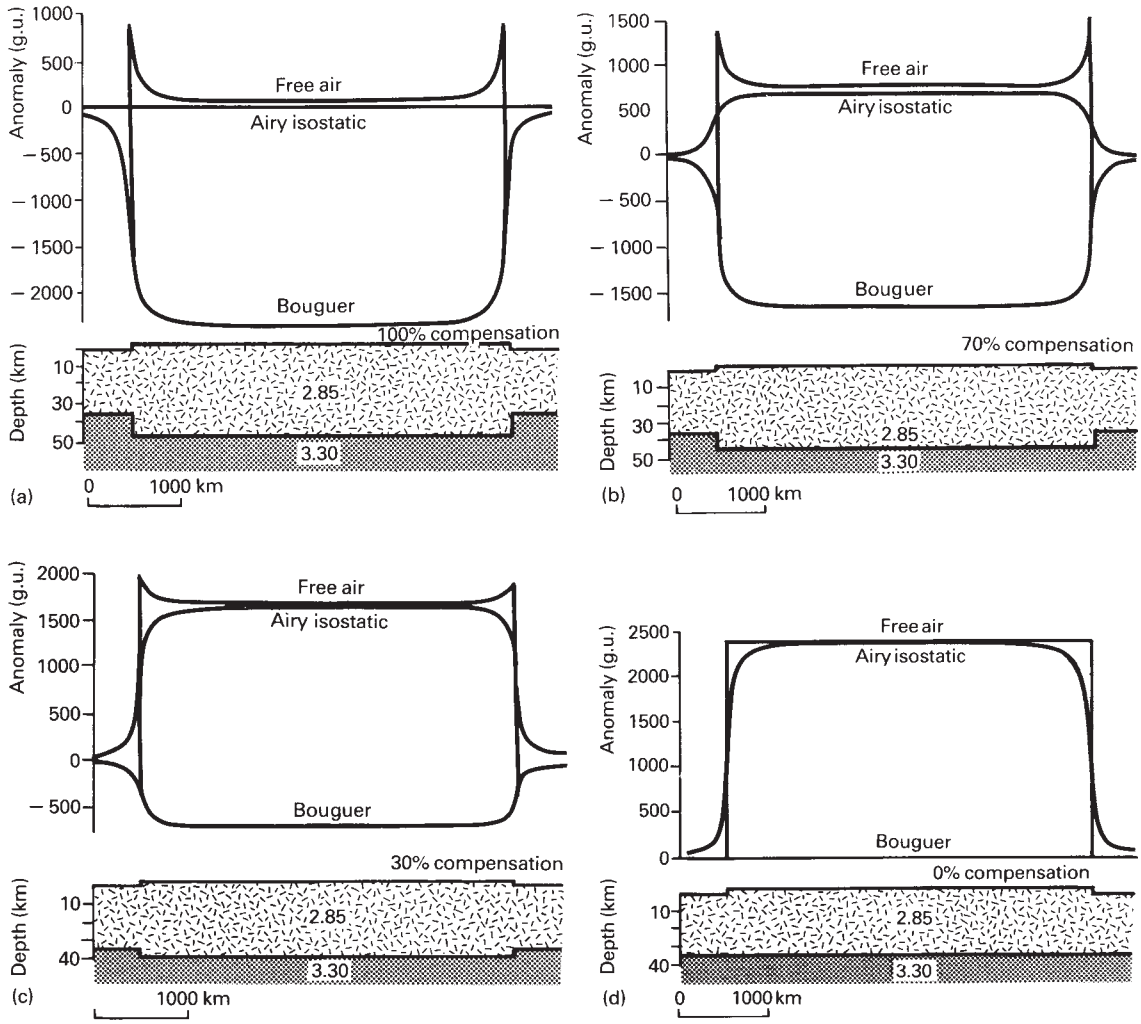


Figure 2.33 *Free air, Bouguer and Airy isostatic anomalies over an idealized mountain range (a) in perfect isostatic equilibrium, (b) with 70% isostatic compensation, (c) with 30% isostatic compensation, (d) uncompensated. Densities in* $Mg\,m^{-3}$.

tion and indicate which type of mechanism is in operation. This is because the compensation occurs at a relatively deep level and the differences in the anomalies produced by a root/antiroot (according to the Airy hypothesis) or by different density units (according to the Pratt hypothesis) would be very small. Moreover, the gravity anomalies over most regions contain short wavelength components resulting from localized, uncompensated geologic structures that obscure the differences in the regional field arising from the different forms of compensation.

A more sophisticated test of isostasy involves the spectral analysis of the topography and gravity anomalies of the region being studied (Watts, 2001). The relationship between gravity and topography changes with wavelength. Moreover, the way in which it changes varies for different isostatic models. Thus by determining the frequency content of the gravity and topographic data it is possible to determine the type of compensation pertaining in the area. The technique also yields an estimate of T_e, the elastic thickness of the lithosphere (Sections 2.11.4, 2.12).

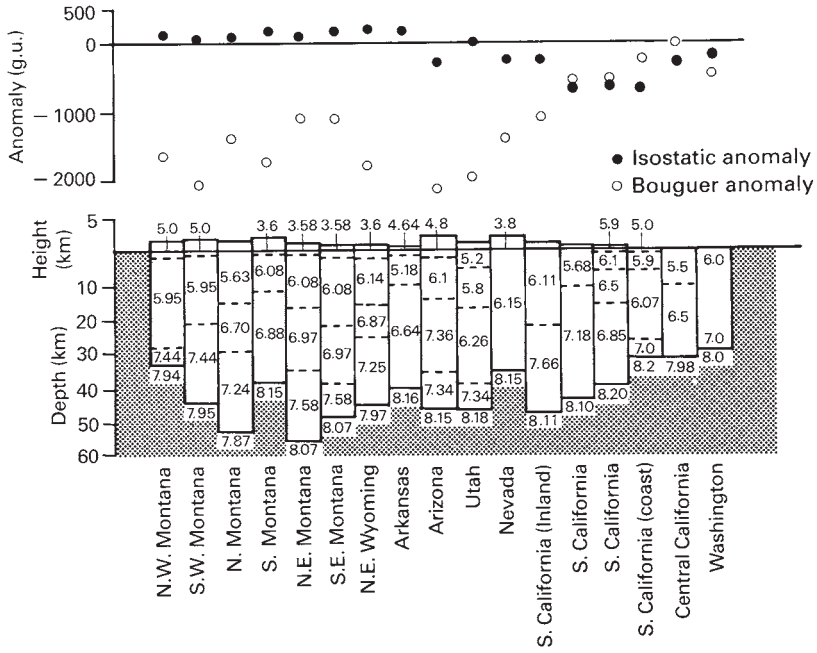


Figure 2.34 *Bouguer and isostatic gravity anomalies and their relation to seismic velocity sections from the western USA. Velocities in km s^{-1} (redrawn from Garland, 1979).*

Information on the geometric form of isostatic compensation can also be gained by a combined analysis of gravity and seismic refraction data, as the latter technique can provide a reasonably detailed picture of the sub-surface structure of the region under consideration. Such studies have demonstrated that the broad isostatic equilibrium of continents and oceans is mainly accomplished by variations in crustal thickness according to the Airy hypothesis. Figure 2.34 shows seismic velocity sections from the western USA in which surface topography is largely compensated by Moho topography, although in several locations density variations in the upper mantle must be invoked to explain the isostatic compensation. A cross-section of the western USA (Fig. 2.35) reveals, however, that crustal thickness is not necessarily related to topographic elevation as the Great Plains, which reach a mean height of 1 km, are underlain by crust 45–50 km thick and the Basin and Range Province, at an average of 1.2 km above mean sea level, is underlain by a crustal thickness averaging 25–30 km (Section 7.3). Clearly, the Basin and Range Province must be partially compensated by a Pratt-type mechanism resulting from the presence of low density material in the upper mantle. Similarly, ocean ridges (Section 6.2) owe their elevation to a region of low density material in the upper mantle rather than to a thickened crust.

There are regions of the Earth's surface that do not conform to the concepts of isostasy discussed here. The hypotheses discussed above all assume that the support of surface features is achieved by their attaining hydrostatic equilibrium with the substrate. In certain areas, however, in particular convergent plate margins, surface features are supported dynamically by horizontal stresses. Such features provide the largest isostatic anomalies observed on the Earth's surface.

2.12 LITHOSPHERE AND ASTHENOSPHERE

It has long been recognized that for large-scale structures to attain isostatic equilibrium, the outermost shell of the Earth must be underlain by a weak layer

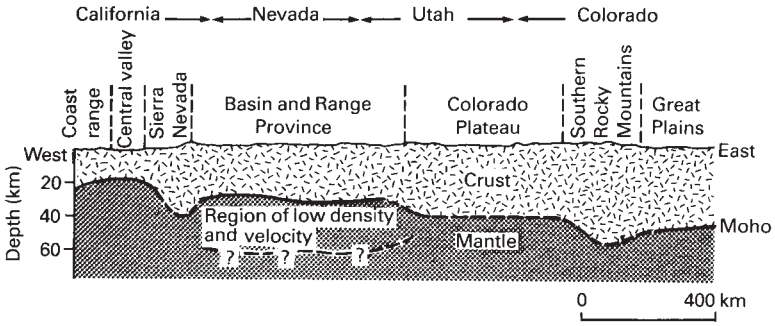


Figure 2.35 *Section from San Francisco, California to Lamar, Colorado based on seismic refraction data (redrawn from Pakiser, 1963, by permission of the American Geophysical Union. Copyright © 1963 American Geophysical Union).*

that deforms by flow. This concept has assumed fundamental importance since it was realized that the subdivisions of the Earth controlling plate tectonic movements must be based on rheology, rather than composition.

The lithosphere is defined as the strong, outermost layer of the Earth that deforms in an essentially elastic manner. It is made up of the crust and uppermost mantle. The lithosphere is underlain by the asthenosphere, which is a much weaker layer and reacts to stress in a fluid manner. The lithosphere is divided into plates, of which the crustal component can be oceanic and/or continental, and the relative movements of plates take place upon the asthenosphere.

However, having made these relatively simple definitions, examination of the several properties that might be expected to characterize these layers reveals that they lead to different ideas of their thickness. The properties considered are thermal, seismic, elastic, seismogenic, and temporal.

Temperature is believed to be the main phenomenon that controls the strength of subsurface material. Hydrostatic pressure increases with depth in an almost linear manner, and so the melting point of rocks also increases with depth. Melting will occur when the temperature curve intersects the melting curve (solidus) for the material present at depth (Fig. 2.36). The asthenosphere is believed to represent the location in the mantle where the melting point is most closely approached. This layer is certainly not completely molten, as it transmits S waves, but it is possible that a small amount of melt is present. The depth at which the asthenosphere occurs depends upon the geothermal gradient and the melting temperature of the mantle materials (Le Pichon

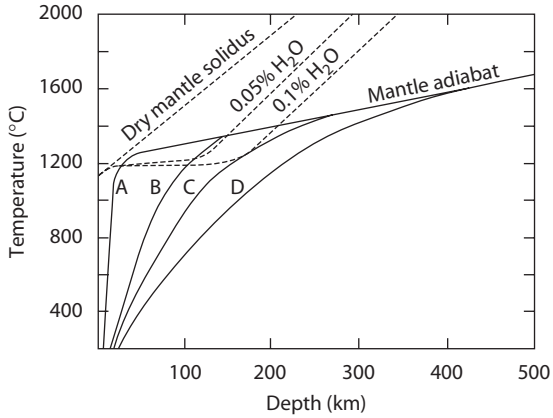


Figure 2.36 *Variation of temperature with depth beneath continental and oceanic regions. A, ocean ridge; B, ocean basin; C, continental platform; D, Archean Shield (redrawn from Condie, 2005b, with permission from Elsevier Academic Press).*

et al., 1973). Beneath ocean ridges, where temperature gradients are high, the asthenosphere must occur at shallow depth. Indeed, since it is actually created in the crestal region (Section 6.10), the lithosphere there is particularly thin. The gradient decreases towards the deep ocean basins, and the lithosphere thickens in this direction, the increase correlating with the depth of water as the lithosphere subsides as a result of contraction on cooling (Section 6.4). The mean lithosphere thickness on this basis beneath oceans is probably 60–70 km. Beneath continents a substantial portion of the observed heat flow is produced within the crust (Section

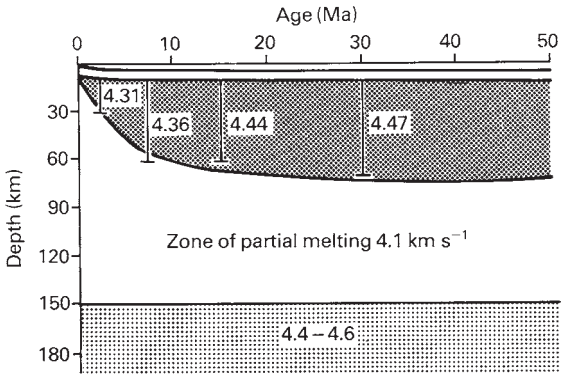


Figure 2.37 *Shear wave model of the thickening of oceanic lithosphere with age. Velocities in km s^{-1} (redrawn from Forsyth, 1975, with permission from Blackwell Publishing). The 150 km transition may be somewhat deeper.*

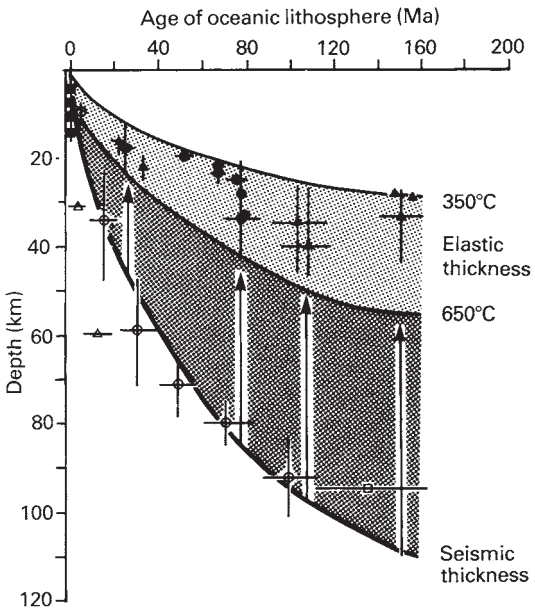


Figure 2.38 *Comparison of short-term "seismic" thickness and long-term "elastic" thickness for oceanic lithosphere of different ages (redrawn from Watts et al., 1980, by permission of the American Geophysical Union. Copyright © 1980 American Geophysical Union).*

2.13), so the temperature gradient in the sub-crustal lithosphere must be considerably lower than in oceanic areas. It is probable that the mantle solidus is not approached until a significantly greater depth, so that the continental lithosphere has a thickness of 100–250 km, being at a maximum beneath cratonic areas (Section 11.3.1).

The depth of the Low Velocity Zone (LVZ) for seismic waves (Section 2.2) agrees quite well with the temperature model of lithosphere and asthenosphere. Beneath oceanic lithosphere, for example, it progressively increases away from the crests of mid-ocean ridges, reaching a depth of approximately 80 km beneath crust 80 Ma in age (Forsyth, 1975) (Fig.2.37). Beneath continents it occurs at greater depths consistent with the lower geothermal gradients (Fig. 2.36). Within the LVZ attenuation of seismic energy, particularly shear wave energy, is very high. Both the low seismic velocities and high attenuation are consistent with the presence of a relatively weak layer at this level. As would be expected for a temperature-controlled boundary, the lithosphere–asthenosphere interface is not sharply defined, and occupies a zone several kilometers thick.

When the Earth's surface is loaded, the lithosphere reacts by downward flexure (Section 2.11.4). Examples include the loading of continental areas by ice sheets or large glacial lakes, the loading of oceanic lithosphere by seamounts, and the loading of the margins of both, at the ocean–continent transition, by large river deltas. The amount of flexure depends on the magnitude of the load and the flexural rigidity of the lithosphere. The latter, in turn, is dependent on the effective elastic thickness of the lithosphere, T_e (Section 2.11.4). Thus, if the magnitude of the load can be calculated and the amount of flexure determined, T_e may be deduced. However as indicated above (Section 2.11.6), T_e may be determined more generally from the spectral analysis of gravity and topographic data. Results obtained by applying this technique to oceanic areas are very consistent. They reveal that the elastic thickness of oceanic lithosphere is invariably less than 40 km and decreases systematically towards oceanic ridges (Watts, 2001) (Fig. 2.38). By contrast, the results obtained for continental areas vary from 5 to 110 km, the highest values being obtained for the oldest areas – the Precambrian cratons. However, McKenzie (2003) maintains that if there are sub-surface density contrasts that have no topographic expression, so-called *buried or hidden* loads, the technique yields an overestimate of the elastic thickness. Such loads are thought to be more common in continental areas, particularly in the cratons, because of their thick and rigid lithosphere. In oceanic areas loads are typically super-

imposed on the crust and expressed in the topography. McKenzie (2003) goes so far as to suggest that, if one makes allowance for buried loads, the elastic thickness of the lithosphere is probably less than 25 km in both oceanic and continental areas. By contrast, Perez-Gussinge & Watts (2005) maintain that T_e is greater than 60 km for continental lithosphere greater than 1.5 Ga in age and less than 30 km for continental areas less than 1.5 Ga in age. They suggest that this is a result of the change in thickness, geothermal gradient, and composition of continental lithosphere with time due to a decrease in mantle temperatures and volatile content (Section 11.3.3). Under tectonically active areas, such as the Basin and Range Province, the elastic thickness may be as small as 4 km (Bechtel *et al.*, 1990). Such very thin elastic thicknesses are undoubtedly due to very high geothermal gradients.

Yet another aspect of the lithosphere is the maximum depth to which the foci of earthquakes occur within it. This so-called *seismogenic thickness* is typically less than 25 km, that is, similar to or somewhat less than the elastic thickness in most areas (Watts & Burov, 2003). On the face of it this appears to lend support to the conclusion of McKenzie (2003) that the spectral analysis of topography and gravity anomalies systematically overestimates T_e, particularly in Precambrian shield areas because of the subdued topography and the presence of buried loads. However, there are alternative explanations that invoke the role of the ductile layer in the lower continental crust in decoupling the elastic upper layer from the lower lithosphere, the role of increased overburden pressure in inhibiting frictional sliding, and the fact that there is some evidence for earthquakes and faulting in the lower crust and upper mantle. It is thought that the latter may occur in the relatively rare instances where the lower crust and/or upper mantle are hydrated (Watts & Burov, 2003).

Thus, the concept of the lithosphere as a layer of uniformly high strength is seen to be over-simplistic when the rheological layering is considered. The upper 20–40 km of the lithosphere are brittle and respond to stress below the yield point by elastic deformation accompanied by transient creep. Beneath the brittle zone is a layer that deforms by plastic flow above a yield point of about 100 MPa. The lowest part, which is continuous with the asthenosphere, deforms by power-law creep and is defined as the region where the temperature increases with depth from 0.55 T_m to 0.85 T_m. The lithosphere is best thought of as a viscoelastic rather than an elastic layer (Walcott, 1970) for, as Walcott demonstrated, the type of deformation experienced depends upon the duration of the applied loads. Over periods of a few thousand years, most of the region exhibiting power-law creep does not deform significantly and consequently is included within the elastic lithosphere. Long term loading, however, occurring over periods of a few million years, permits power-law deformation to occur so that this region then belongs to the asthenosphere.

The lithosphere can, therefore, be defined in a number of different ways that provide different estimates of its thickness. This must be borne in mind throughout any consideration of plate tectonic processes.

The asthenosphere is believed to extend to a depth of about 700 km. The properties of the underlying region are only poorly known. Seismic waves that cross this region do not suffer great attenuation (Section 9.4), and so it is generally accepted that this is a layer of higher strength, termed the *mesosphere*. The compositional and rheological layering of the Earth are compared in Fig. 2.39.

2.13 TERRESTRIAL HEAT FLOW

The study of thermal processes within the Earth is somewhat speculative because the interpretation of the distribution of heat sources and the mechanisms of heat transfer are based on measurements made at or near the surface. Such a study is important, however, as the process of heat escape from the Earth's interior is the direct or indirect cause of most tectonic and igneous activity.

The vast majority of the heat affecting the Earth's surface comes from the Sun, which accounts for some 99.98% of the Earth's surface energy budget. Most of this thermal energy, however, is reradiated into space, while the rest penetrates only a few hundred meters below the surface. Solar energy consequently has a negligible effect on thermal processes occurring in the interior of the Earth. The geothermal energy loss from heat sources within the Earth constitutes about 0.022% of its surface energy budget. Other sources of energy include the energy generated by the gradual deceleration of the

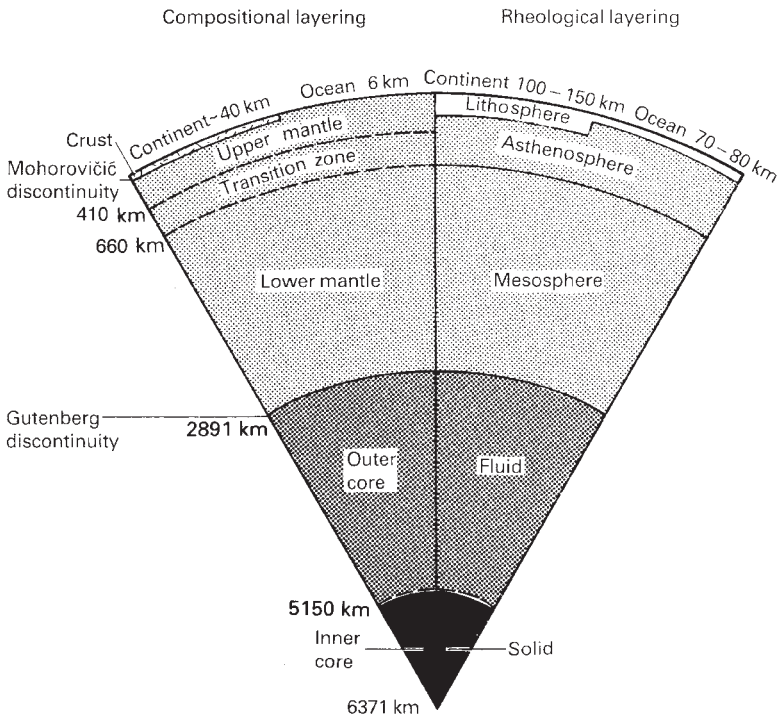


Figure 2.39 *Comparison of the compositional and rheological layering of the Earth.*

Earth's rotation and the energy released by earthquakes, but these make up only about 0.002% of the energy budget. It is thus apparent that geothermal energy is the major source of the energy which drives the Earth's internal processes.

It is believed that the geothermal energy is derived in part from the energy given off during the radioactive decay of long-lived isotopes, in particular K^{40}, U^{235}, U^{238}, and Th^{232}, and also from the heat released during the early stages of the formation of the Earth. These isotopes would account for the present geothermal loss if present in proportions similar to those of chondritic meteorites. Radioactive decay is exponential, so that during the early history of the Earth the concentration of radioactive isotopes would have been significantly higher than at present and the thermal energy available to power its internal processes would have been much greater (Section 12.2). Currently accepted models for the formation of the Earth require an early phase of melting and differentiation of its originally homogeneous structure. This melting is believed to have been powered in part by thermal energy provided by the decay of short-lived radioactive isotopes such as Al^{26}, Fe^{60}, and Cl^{36}. The differentiation of the Earth would also have contributed energy to the Earth arising from the loss in gravitational potential energy as the dense iron-nickel core segregated to a lower energy state at the center of the Earth.

The heat flow through a unit area of the Earth's surface, H, is given by:

$$H = K\frac{\delta T}{\delta z}$$

where $\delta T/\delta z$ is the thermal gradient perpendicular to the surface and K the thermal conductivity of the medium through which the heat is flowing. The units of H are $mW\,m^{-2}$.

On land, heat flow measurements are normally made in boreholes. Mercury maximum thermometers or thermistor probes are used to determine the vertical temperature gradient. Thermal conductivity is measured on samples of the core using a technique similar to the Lee's disc method. Although appearing relatively simple, accurate heat flow measurements on land are

difficult to accomplish. The drilling of a borehole necessitates the use of fluid lubricants that disturb the thermal regime of the borehole so that it has to be left for several months to allow the disturbance to dissipate. Porous strata have to be avoided as pore water acts as a heat sink and distorts the normal thermal gradients. Consequently, it is rarely possible to utilize boreholes sunk for the purposes of hydrocarbon or hydrogeologic exploration. In many areas readings may only be undertaken at depths below about 200 m so as to avoid the transient thermal effects of glaciations.

Heat flow measurements are considerably easier to accomplish at sea. The bottom temperatures in the oceans remain essentially constant and so no complications arise because of transient thermal perturbations. A temperature probe is dropped into the upper soft sediment layer of the seabed and, after a few minutes' stabilization, the temperature gradient is measured by a series of thermistor probes. A corer associated with the probe collects a sediment sample for thermal conductivity measurements; alternatively, the role of one of the thermistors can be changed to provide a source of heat. The change in the temperature of this probe with time depends on the rate at which heat is conducted away from it, and this enables a direct, *in situ* measurement of the thermal conductivity of the sediment to be made.

A large proportion of geothermal energy escapes from the surface by conduction through the solid Earth. In the region of the oceanic ridge system, however, the circulation of seawater plays a major role in transporting heat to the surface and about 25% of the geothermal energy flux at the Earth's surface is lost in this way.

The pattern of heat flow provinces on the Earth's surface broadly correlates with major physiographic and geologic subdivisions. On continents the magnitude of heat flow generally decreases from the time of the last major tectonic event (Sclater *et al.*, 1980). Heat flow values are thus low over the Precambrian shields and much higher over regions affected by Cenozoic orogenesis. Within the oceans the heat flow decreases with the age of the lithosphere (Section 6.5), with high values over the oceanic ridge system and active marginal seas and low values over the deep ocean basins and inactive marginal seas.

The average heat flux in continental areas is 65 mW m^{-2}, and in oceanic areas 101 mW m^{-2}, of which about 30% is contributed by hydrothermal activity at the mid-oceanic ridge system (Pollack *et al.*, 1993). As 60% of the Earth's surface is underlain by oceanic crust, about 70% of the geothermal energy is lost through oceanic crust, and 30% through continental crust.

FURTHER READING

Anderson, D.L. (2007) *New Theory of the Earth*, 2nd edn. Cambridge University Press, Cambridge, UK.

Bott, M.H.P. (1982) *The Interior of the Earth, its Structure, Constitution and Evolution*, 2nd edn. Edward Arnold, London.

Condie, K.C. (2005) *Earth as an Evolving Planetary System*. Elsevier, Amsterdam.

Fowler, C.M.R. (2005) *The Solid Earth: an introduction to global geophysics*, 2nd edn. Cambridge University Press, Cambridge.

Jacobs, J.A. (1991) *The Deep Interior of the Earth*. Chapman & Hall, London.

Nicolas, A. (1989) *Structure of Ophiolites and Dynamics of Oceanic Lithosphere*. Kluwer Academic Publishers, Dordrecht.

Park, R.G. (1988) *Geological Structures and Moving Plates*. Blackie, London and Glasgow.

Ranalli, G. (1995) *Rheology of the Earth*, 2nd edn. Chapman & Hall, London.

Stein, S. & Wysession, M. (2003) *An Introduction to Seismology, Earthquakes, and Earth Structure*. Blackwell Publishing, Oxford.

Twiss, R.J. & Moores, E.M. (2006) *Structural Geology*, 2nd edn. W.H. Freeman, New York.