Mantle Convection in the Earth and Planets

GERALD SCHUBERT

University of California, Los Angeles

DONALD L. TURCOTTE

Cornell University

PETER OLSON

The Johns Hopkins University

CAMBRIDGE
UNIVERSITY PRESS

© Cambridge University Press 2004

First published in printed format 2001

3

Structure and Composition of the Mantle

3.1 Introduction

In this chapter we review what is currently known about the structure and the bulk composition of the mantle, with emphasis on how these may influence the style of convection, and how they may be influenced by it. Brief descriptions of the crust and the core are also included, again with an emphasis on those aspects of each region most germane to mantle convection. There are several monographs dealing with Earth structure and composition in a general context, e.g., Ringwood (1975, 1979) and Anderson (1989).

The study of mantle structure is by tradition the province of seismology, while mantle composition has historically been a subject for high-pressure and high-temperature mineralogy, petrology, and geochemistry. A great many important advances have recently been made in these areas, with the result that the studies of mantle composition, structure, and dynamics are now closely related. New findings from seismology, mineral physics, and isotope geochemistry are quickly applied as constraints on models of mantle dynamics. There are also interactions in the other direction. Each new step in understanding the physics of convection is quickly incorporated into new interpretations of mantle structure. This interdisciplinary style of research is perhaps the single most important reason for the emerging view of the deep Earth as a unified physical and chemical system.

3.2 Spherically Averaged Earth Structure

The determination of elastic parameters and density throughout the Earth using observations of seismic waves and other constraints is the prototype inverse problem in geophysics. Like many inverse problems, it is formally nonunique and suffers from practical difficulties such as incomplete sampling and errors in the data. In spite of this, it is remarkable how much is now known about mantle structure, and this is particularly true of spherically averaged properties.

The major divisions of the Earth's interior are shown in Figure 3.1. The crust, mantle, and core were recognized in the early part of this century following Mohorovicic's 1909 discovery of the crust–mantle boundary (Moho) and Gutenberg's (1913) determination of the outer core radius. Lehmann (1936) inferred the existence of the inner core in 1936 and by 1939 Jeffreys (1939) had produced compressional and shear wave velocity profiles featuring a transition zone between about 400 km and 1,000 km depth. Thus, a nearly complete picture of the first-order spherical structure of the Earth was obtained prior to 1940. This period
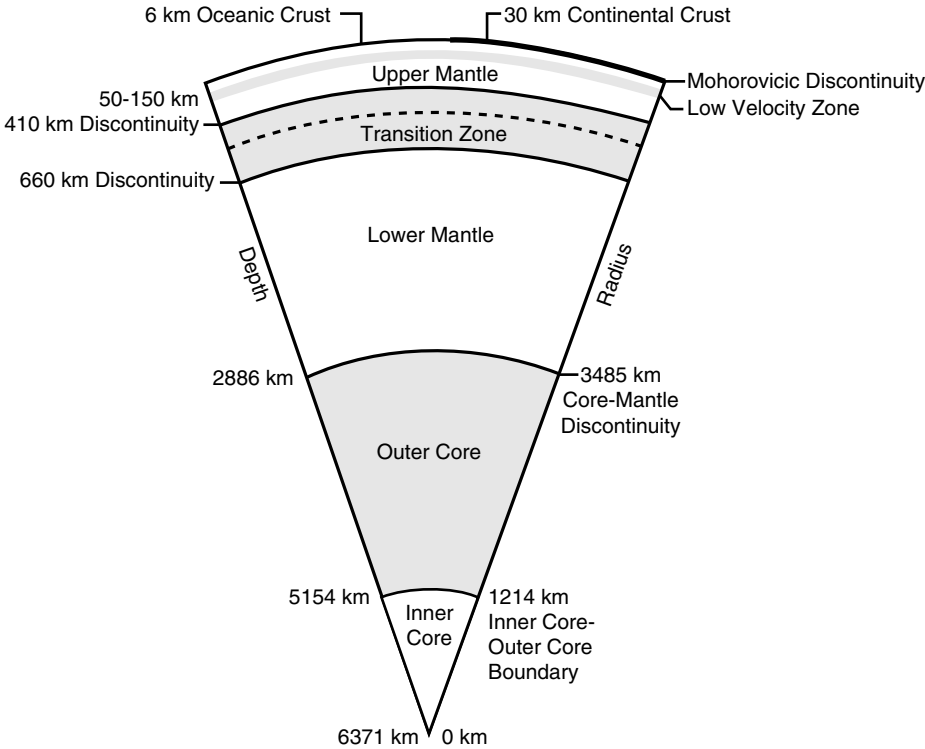
Figure 3.1. The major components of the Earth's interior are the crust, the mantle, and the core. The oceanic crust has a basaltic composition and a mean thickness of about 6 km. The continental crust has a more silicic composition and a mean thickness of about 30 km. The mantle has an ultrabasic composition; the compositional boundary between the crust and mantle is the Mohorovicic seismic discontinuity (the Moho). The mantle has major seismic discontinuities at depths of about 410 and 660 km. The core is primarily iron; the outer core is liquid and the inner core is solid. The depths and radii of major boundaries are shown assuming spherical symmetry.

culminated with the publication of standard travel time curves for major seismic body wave phases, the JB tables (Jeffreys and Bullen, 1940), which are still in use today.

A method for obtaining radial density variations from radial profiles of compressional and shear wave velocities was developed by Williamson and Adams (1923). It provides a simple equation of state for describing compressibility in chemically homogeneous layers. Within such a layer, the radial variation in density ρ can be expressed in terms of any other two state variables. In particular we consider the pressure p and the entropy s and write

$$\frac{d\rho}{dr} = \left(\frac{\partial \rho}{\partial p}\right)_s \frac{dp}{dr} + \left(\frac{\partial \rho}{\partial s}\right)_p \frac{ds}{dr} \tag{3.2.1}$$

where r is the radial coordinate and the subscripts s and p refer to isentropic and isobaric variations, respectively. An isentropic process is a reversible process in which there is no heat transfer (adiabatic) and an isobaric process is a process at constant pressure. Applications of thermodynamics to the mantle will be discussed in greater detail in Section 6.8.

The thermodynamic derivative on the right side of (3.2.1), $(\partial\rho/\partial p)_s$, is related to the adiabatic compressibility χ_a of a material. The definition of χ_a is

$$\chi_a \equiv \frac{1}{\rho}\left(\frac{\partial\rho}{\partial p}\right)_s = \frac{1}{K_a} \tag{3.2.2}$$

where K_a is the adiabatic bulk modulus. The connection with elasticity comes from the relation between K_a and the velocities of the seismic compressional and shear waves, V_P and V_S, respectively, where

$$V_P = \left(\frac{K_a + 4\mu/3}{\rho}\right)^{1/2} \tag{3.2.3}$$

$$V_S = \left(\frac{\mu}{\rho}\right)^{1/2} \tag{3.2.4}$$

and μ is the shear modulus or rigidity of the solid. Elimination of the shear modulus from (3.2.3) and (3.2.4) gives

$$\Phi \equiv V_P^2 - \frac{4}{3}V_S^2 = \frac{K_a}{\rho} \tag{3.2.5}$$

where Φ is the seismic parameter. From (3.2.2) and (3.2.5), we can write the thermodynamic derivative $(\partial\rho/\partial p)_s$ simply as

$$\left(\frac{\partial\rho}{\partial p}\right)_s = \frac{1}{\Phi} \tag{3.2.6}$$

The radial profiles of V_P and V_S determined from seismology also give $\Phi(r)$ and $(\partial\rho/\partial p)_s$ as a function of r.

For a homogeneous layer that is well mixed, e.g., by convection, it is appropriate to assume that the layer is isentropic and $ds/dr = 0$. With this assumption and (3.2.6), the variation of density in (3.2.1) can be written as

$$\frac{d\rho}{dr} = \frac{1}{\Phi}\frac{dp}{dr} \tag{3.2.7}$$

The radial pressure derivative dp/dr in (3.2.7) is given to a good approximation throughout the mantle by the hydrostatic equation

$$\frac{dp}{dr} = -\rho g \tag{3.2.8}$$

where g is the acceleration of gravity. Substitution of (3.2.8) into (3.2.7) gives the Adams–Williamson equation for the variation of density with radius in the mantle:

$$\left(\frac{d\rho}{dr}\right)_s = \frac{-\rho g}{\Phi} \tag{3.2.9}$$

The subscript s on the left of (3.2.9) indicates that the process is isentropic. It should be emphasized that this result is valid only if the composition is uniform.

The acceleration of gravity g in the mantle also varies with radius. For a spherically symmetric Earth model $g(r)$ satisfies the Poisson equation

$$\frac{1}{r^2}\frac{d}{dr}\left(r^2 g\right) = 4\pi G\rho \tag{3.2.10}$$

where G is the universal constant of gravitation. Integration of (3.2.9) and (3.2.10), with $\Phi(r)$ known from seismology, gives a spherical Earth model, consisting of the radial variation of spherically averaged density and gravity. Integration of these coupled first-order differential equations requires two boundary conditions or constraints which are provided by the Earth's mass and moment of inertia; the densities of crustal and upper mantle rocks are also used to fix the densities at the tops of individual spherical layers. With $\rho(r)$ and $g(r)$ determined, the variation with radius of the adiabatic bulk modulus and rigidity can readily be found from (3.2.5) and (3.2.4), for example. The spherical Earth model also consists of $\mu(r)$ and $K_a(r)$.

Bullen (1936, 1940) first used the above procedure to obtain a six-layer Earth model, consisting of the crust (layer A), from the Earth's surface to the Moho (at a mean depth of 6 km beneath the oceans and 30 km beneath the continents), an adiabatic upper mantle (layer B), from the Moho to a depth of 400 km, an adiabatic lower mantle (layer D), from a depth of 1,000 km to a depth of 2,900 km, an adiabatic outer core (layer E), from a depth of 2,900 km to 5,100 km, and an adiabatic inner core (layer F), from 5,100 km depth to the center of the Earth. He found that the adiabatic approximation was not appropriate for the transition zone (layer C), from a depth of 400 km to 1,000 km, and instead used a polynomial function of radius to represent the density variation in this layer.

The resulting model A, as it was called, was an immediate success. Later, Bullen published model B, in which he added the assumption of continuity in the bulk modulus and its pressure derivative across the core–mantle boundary. This allowed him to further subdivide the lower mantle into layers D′, to a depth of 2,700 km, and D″, between 2,700 km and 2,900 km depth. Bullen's models A and B are masterpieces of inductive science. They have been superseded by Earth models derived from larger data sets, and among all of his layer notation, only D″ has survived. However, these models were remarkably accurate. Perhaps more importantly, they introduced a new subject into geophysics – the use of spherical Earth models to infer mantle and core composition.

Modern Earth models are based on vastly larger seismic body wave data sets than were available to Bullen, and we now have normal mode frequencies from both toroidal oscillations (the mode equivalent of seismic Love waves) and spheroidal oscillations (equivalent to seismic Rayleigh waves) of the whole Earth. Normal mode frequencies are sensitive to the density distribution as well as the distribution of V_P and V_S, and incorporating them into Earth models substantially tightens the constraints on density, particularly in the lower mantle and outer core. The mathematics of the inversion procedure can be found in a number of review papers and texts, including Gilbert and Dziewonski (1975), Aki and Richards (1980), and Dziewonski and Anderson (1981). Spherically symmetric Earth models such as Dziewonski and Anderson's 1981 Preliminary Reference Earth Model (PREM) are derived using 2×10^6 P-wave (compressional wave) and 2×10^5 S-wave (shear wave) travel times and approximately 10^3 normal mode frequencies. The technique used in constructing these models is to apply body wave data to resolve fine structure and normal modes to determine the average density in each layer of the model. Uncertainties in density are quite low for averages over a finite depth interval. In the lower mantle and outer core, the uncertainty is less than 2% for averaging intervals of 400 km or more (Masters, 1979). This is somewhat

misleading, however, because the uncertainty increases rapidly as the averaging interval is reduced. In particular, the spherical Earth models are not able to accurately resolve discontinuities. The procedure for treating seismic and density discontinuities is to prescribe their depth a priori and allow the inversion procedure to determine the best-fitting values for properties in the layers on either side. Because precision decreases as the averaging interval decreases, it is difficult to resolve accurately the fine structure in important regions like the transition zone, where there are several closely spaced discontinuities. This limitation should always be kept in mind when interpreting the fine structure in spherical Earth models.

Profiles of V_P, V_S, and density ρ from model PREM are shown in Figure 3.2 for the whole Earth, and in detail for the upper mantle in Figure 3.3. The variations in bulk modulus
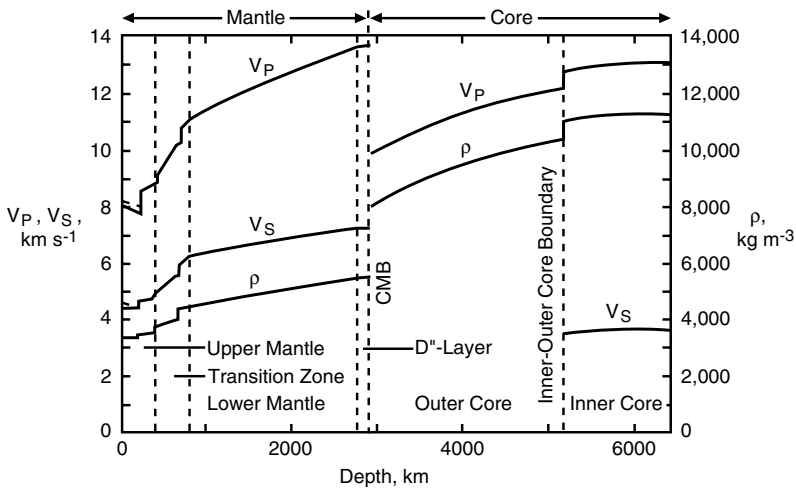

Figure 3.2. Spherical Earth model PREM (Dziewonski and Anderson, 1981). The seismic velocities V_P, V_S and density ρ are given as a function of depth.
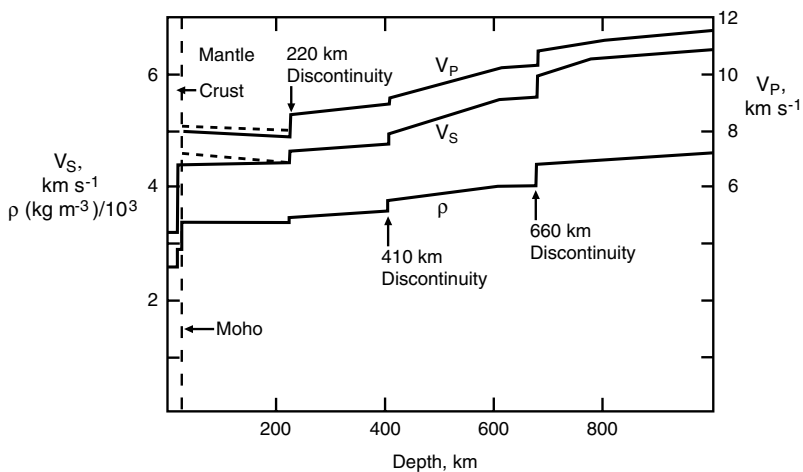

Figure 3.3. Structure of the upper mantle from the spherical Earth model PREM (Dziewonski and Anderson, 1981).
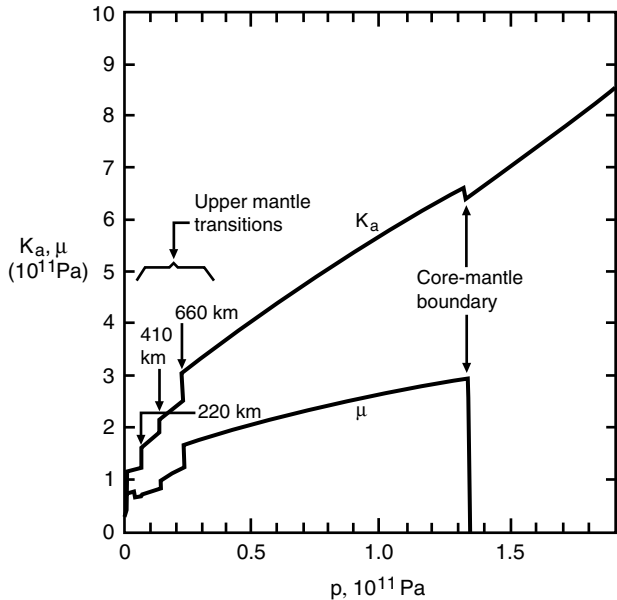
Figure 3.4. Dependences of the Earth's adiabatic bulk modulus K_a and shear modulus or rigidity μ on pressure in the spherical Earth model PREM (Dziewonski and Anderson, 1981).

and rigidity with pressure through the mantle are shown in Figure 3.4. PREM is constructed assuming three first-order discontinuities in the upper mantle, at depths of 220 km, 400 km, and 650 km. Like its predecessor PEM (Parametric Earth Model, Dziewonski et al., 1975), the profiles consist of polynomials within each layer. PREM assumes isotropy everywhere except between the Moho and 220 km depth, where transverse anisotropy is permitted. PREM also contains anelasticity, through the quality factor Q (inverse attenuation), and frequency-dependent elastic moduli. A major improvement over earlier models, such as Bullen's, is the presence of fine structure. Despite the fact that fine structure is the least certain component, it is the main feature that distinguishes one spherical model from another. Small differences in fine structure can lead to different interpretations regarding composition and its variation with depth in the mantle, which in turn may lead to grossly different inferences about the style of convection. Modern spherical Earth models (e.g., IASPEI 1991, see Kennett, 1991) have verified the major subdivisions listed in Table 3.1. These divisions provide the basis for discussions of the Earth's composition.

3.3 The Crust

In terms of its mass, the crust is a negligible part of the Earth, comprising less than 0.5% of the total (Table 3.1). But the significance of the crust for mantle convection is far greater than its fraction of the Earth's mass. The crust is the end product of mantle differentiation. Since differentiation of the mantle occurs primarily by partial melting, which is itself a product of convection, the crust yields information on the history of convection which is not obtainable by any other means. The crust also contains the most extreme lateral heterogeneities in the Earth, and this has a direct influence on the style of mantle convection. The continental

Table 3.1. *Major Divisions of the Earth*

Region	Depth Range (km)	Mass (10^{21} km)	Mass Fraction (%)	Average Density (kg m^{-3})
Oceanic crust	0–6	6	0.1	3,000
Continental crust	0–30	19	0.3	2,700
Upper mantle	(6,30)–410	615	10.3	3,350
Transition zone	410–660	415	7.0	3,860
Lower mantle	660–2,886	2,955	49.6	4,870
Outer core	2,886–5,140	1,867	31.1	11,000
Inner core	5,140–6,371	98	1.6	12,950
Whole Earth	0–6, 371	5,975	100	5,515

crust contains a substantial fraction of the Earth's incompatible elements including the heat-producing elements uranium, thorium, and potassium. This fractionation strongly influences the strength of mantle convection.

The crust (continental and oceanic) is bimodal in nearly every respect: thickness, age, density, and composition. It is not possible to overemphasize this distinction, particularly with regard to the role of the crust in mantle convection. On average, the continental crust is old (mean age $\simeq$ 2,000 Myr) while the oceanic crust is young (mean age $\simeq$ 80 Myr). Oceanic crust is created continuously at spreading centers, the present rate of creation being 2.8 km^2 yr^{-1} (volumetrically about 17 km^3 yr^{-1}), and is subducted back into the mantle at approximately the same rate. It is formed, almost entirely, by a single process – partial melting of an upper mantle composed of olivine, pyroxene, and garnet – resulting in a remarkably homogeneous product, mid-ocean ridge basalt (MORB). In contrast, most of the continental crust has resided on Earth's surface long enough to have experienced a sequence of collision events, intrusions of melts from below, episodes of loading and unloading, infiltration of metamorphic fluids, erosion and alterations in its thermal regime. One of the primary factors that accounts for the complexities of the continental crust and the relative simplicity of oceanic crust is the great disparity in ages. However, the silicic rocks of the continental crust are much weaker than the mafic rocks of the oceanic crust. Thus intraplate deformation is largely restricted to the continents as indicated by the distribution of earthquakes given in Figure 2.2. The relative weakness of continental rocks may have important implications for the tectonic evolution of the Earth and other planets. The history of plate tectonics requires the internal deformation of plates (Dewey, 1975); the required deformation occurs within the continents. As will be discussed in Chapter 14, the absence of "soft" continents on Venus may be the reason that Venus does not have plate tectonics.

3.3.1 Oceanic Crust

The same sequence of structural and lithologic units characterizes the oceanic crust on a global scale. Figure 3.5 shows the major layers in a typical cross-section, and Table 3.2 summarizes their properties. This stratigraphy was first identified from reflection and refraction seismic profiles; the connection between the seismic structure and the petrologic units was established through comparisons with ophiolite sequences, believed to be sections of oceanic crust and upper mantle thrust onto the continental crust. There are a number of excellent reviews of this subject, for example, Orcutt et al. (1984), and Purdy and Ewing
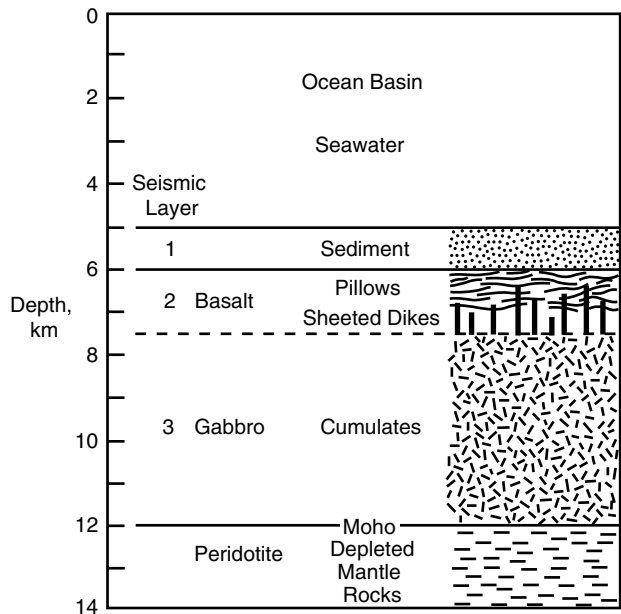
Figure 3.5. Idealized cross-section of the oceanic crust. Layer 1 is composed of sediments which have been deposited on the igneous layers 2 and 3. Layers 2 and 3 have essentially the same basaltic composition. Layer 2 is made up of the extrusive pillow basalts and the intrusive sheeted dike complex. Both have a fine-grained basaltic structure due to rapid cooling and have been extensively altered by hydrothermal flows. Layer 3 is made up of coarse-grained gabbros which cooled slowly in the magma chambers in the crust at the ridge axis.

Table 3.2. *Oceanic Crust Structure*

Seismic Layer	Average Thickness (km)	V_P (km s^{-1})	Average Density (kg m^{-3})
Seawater	4.6	1.5	1,026
Sediments	Variable	2.1 ± 0.7	2,000
Basaltic	1.6 ± 0.7	5.1 ± 0.6	3,000
Gabbroic	4.7 ± 1.4	6.8 ± 0.2	3,000
Upper mantle	–	8.15 ± 0.2	3,350

(1986). There are three important lithologic units in ocean crust between the water column (4.6 km average depth) and the upper mantle. Layer 1 is deep-sea sediments, about 0.5 km thick on average, but highly variable, with a statistical increase in thickness with increasing crustal age. The sediments are a fine-grained mixture of clays and various oozes, except where contaminated by terrigenous sediments delivered to the deep ocean by large river systems. Poorly consolidated, and with a high pore water content, their average V_P velocity and density are only 2.1 km s^{-1} and 2,000 kg m^{-3}, respectively. Layer 2 is composed of fine-grained tholeiitic basalts and basaltic glasses. The upper part of layer 2 is composed of pillow basalts, which are the products of basaltic magma extrusions onto the seafloor and their interactions with seawater. The lower part of layer 2 is made up of the coarser-grained basaltic sheeted dike complex. During each eruption the basaltic magma flows from the magma chamber at depth to the surface through a vertical crack or dike. The solidification

of the magmas in these dikes produces the sheeted dike complex. The average thickness of layer 2 is 1.6 km and it is considered to be the major source of seafloor magnetic anomalies (Smith and Banerjee, 1986). It is also seismically anisotropic, with the fast axis perpendicular to the direction of seafloor spreading where the crust was formed. This property has been ascribed to cracks and fissures oriented perpendicular to the original spreading direction, acquired during cooling at the ridge crest (Stephen, 1985). Layer 3 contains coarse-grained gabbros and related cumulates with essentially the same basaltic composition as layer 2. The coarse-grained structure is the result of slow cooling in the magma chamber at the base of the ocean crust at the ridge crest. Layer 3 averages 4.7 km in thickness.

The complete oceanic crust is typically 6–7 km thick. There is no systematic variation in thickness of layers 2 and 3 with age, in accord with the source of MORB being localized at spreading centers. Oceanic crust can range in thickness from 2 to 37 km; the partitioning of the crust into layer 2 and layer 3 characterizes even the thickest oceanic crust (Mutter and Mutter, 1993). Variations in the total thickness of oceanic crust result primarily from changes in the thickness of the lower oceanic crust, Layer 3 (Mutter and Mutter, 1993). The greatest thickness variations in oceanic crust occur along hot spot tracks, and at rifted margins where hot spots appear to have been initiated (White and McKenzie, 1989). The significance of these facts for the mantle plume theory of hot spot formation will be discussed in Chapter 11.

Numerous dredge samples have demonstrated the remarkable uniformity of the MORB bulk composition (Basaltic Volcanism Study Project, 1981). MORB is a tholeiitic basalt, slightly oversaturated in silica (see Table 2.2) with a relatively simple mineralogy – labradorite, calcic pyroxene, and minor amounts of olivine. The density of MORB lies between 2,900 and 3,000 $kg\,m^{-3}$. Chemical analyses show a high concentration of Al_2O_3 and high Na/K ratios, relative to continental basalts (Hall, 1987). The geochemistry of MORB indicates it is derived from mantle material that is depleted, relative to chondritic meteorites, in incompatible elements such as K, Rb, Sr, U, Th, and in the rare earth elements. This pattern indicates that the parent mantle material is not "primitive," but rather it has been differentiated, and has thereby lost some of its original inventory of low melting point components. It is generally accepted that the continental crust is the major complementary reservoir for the depleted mantle (O'Nions and Pankhurst, 1974; Zindler and Hart, 1986). The degree of differentiation of MORB parent material, its relationship to the continental crust, and the fraction of the mantle involved in producing MORB all bear on the question of the structure of mantle convection, and will be treated more fully in Chapter 12.

3.3.2   Continental Crust

The continental crust is distinctive from the oceanic crust in many ways. Its bulk composition is more silicic and this results in its bulk density being less than that of the oceanic crust. This low density and its greater thickness provide a gravitational stability to the continental lithosphere. Thus, continental crust is not readily recycled into the Earth's interior as the oceanic crust is at subduction zones. As a result the mean age of the continental crust is more than one order of magnitude greater than the mean age of the oceanic crust. The continental crust is made up of igneous, metamorphic, and sedimentary components. Some of the igneous rocks have a basaltic composition similar to that of the oceanic crust, but most of them have a more silicic composition. Examples of silicic volcanism are the intrusive and extrusive volcanics of Andean-type subduction zones. A long-standing debate in geology concerns the origin of these granitic volcanics. Are they products of the mantle or of crustal fractionation? The current consensus favors crustal fractionation.

The hydrologic cycle plays an essential role in the evolution of the continents. Mountain ranges erode to mean sea level in relatively short geological times ($\approx 50\,\text{Myr}$) generating sediments that collect on the continents, in internal sedimentary basins, and on the ocean floor. Some of the sediments on the ocean floor are recycled into the interior of the Earth at subduction zones but a substantial fraction are added to the continents in accretional prisms. Sediment piles can have thicknesses of 10 km or more and the high temperatures and pressures at the base of the piles can result in chemical compositional changes. These become the metamorphic rocks of the continental crust.

Unlike the oceanic crust, which has a single dominant mode of origin and a nearly universal stratigraphy, the continental crust is a heterogeneous aggregation of blocks, each having a very individual structure and history. It is somewhat misleading to discuss the continental crust in terms of its "average" properties, since it is possible that such an average may not occur. With this reservation in mind, we show in Figure 3.6 a sequence of representative seismic cross-sections through hypothetical examples of continental crust. The mean depth to Moho is about 35 km, and varies from 20 km or less beneath the continental slope at passive margins to 70–80 km beneath collision zones such as the Himalayas.

Until recently the structure of the deep continental crust was relatively unexplored, and this is reflected in the shortage of quantitative information on the deep crust in the older literature. This situation is rapidly changing, with organized exploration programs on every continent producing reflection and refraction data. In addition to an extensive survey literature on the continental crust as a whole (Burchfiel, 1983; Taylor and McLennan, 1985; Meissner, 1986; Windley, 1995), there are some interesting models of the lower crust (Fuchs, 1986; Mooney and Brocher, 1987) motivated by findings from deep seismic profiling experiments (Barazangi and Brown, 1986; Matthews and Smith, 1987).

Distinct layering can be found in the continental crust but it is more variable than in the oceanic crust. Much of the continental crust is covered with a thin veneer of sediments. The boundary between these sediments and the igneous and metamorphic rocks of the upper continental crust is known as basement. The basement rocks are exposed principally in the Archean shields and have an overall composition roughly equivalent to a granodiorite. The continental crust is generally divided into an upper crust and a lower crust. The upper
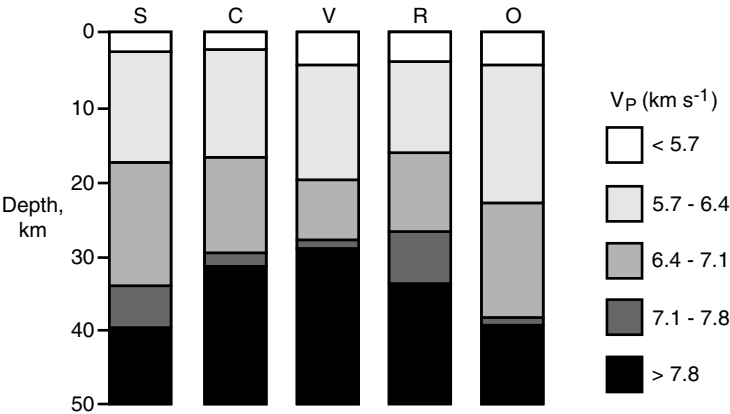

Figure 3.6. Idealized cross-sections of continental crust, showing the variations of V_P with depth. S = shields, C, V = Precambrian and Paleozoic platforms, R = rift zones, O = Mesozoic and Cenozoic orogenic belts. After Meissner (1986).

crust includes both the sediments and the crystalline basement rocks. Typical thicknesses of the upper crust are in the range 20–25 km and the basement rocks have seismic velocities V_P ranging from 5.7 to 6.3 km s^{-1}. The lower crust has thicknesses in the range 15–20 km and seismic velocities V_P in the range 6.3–7.5 km s^{-1} (Rudnick and Fountain, 1995). The boundary between upper and lower crust is often indistinct; where it is observed it sometimes coincides with the Conrad seismic discontinuity, a feature in many older models of deep crustal structure. Depth profiles of seismic velocities, including the Conrad seismic discontinuity, have been inferred from the refraction of earthquake-generated seismic waves.

Deep seismic reflection profiling studies have provided additional information on the structure of the continental crust (Brown, 1991). Deep crustal reflection studies have demonstrated that the two-layer model of the continental crust is generally applicable, and furthermore it now appears that the structural contrast between the layers mirrors a contrast in rheology, which has important implications for continental tectonics. In stable shield and platform regions, Archean age crust is rather transparent to high-frequency seismic waves, particularly in the lower crust, suggesting the absence of sharp gradients in composition. Indeed, the Moho beneath shields and platforms is often not observed in reflection profiles. Regions that have experienced more recent deformation have a much different reflectivity. The upper crust in these regions is also relatively transparent to seismic reflections, a property attributed to the general lack of horizontal layering. This is an indication of brittle deformation in response to long-term tectonic forces and a general absence of pervasive ductile flow. Fine structure is generally oriented vertically, and probably consists of dikes, vertical shear zones, and plug-like intrusions. In contrast, the lower Phanerozoic crust is highly reflective at seismic exploration wavelengths. The most probable explanation is the presence of many fine-scale (100 m thick) horizontal laminations of various rock types. This is an indication of pervasive ductile deformation. The lower crust, composed of granulite facies metamorphic rocks, is ductile and generally more deformable than the granitic upper crust, and it is also more deformable than the underlying mantle. The resulting "sandwich" rheological structure may explain why continental deformation is generally more diffuse than deformation in the oceanic lithosphere. In particular, the weak lower crust probably accounts for the thin-skin style of deformation, including low-angle thrust faults that characterize continental tectonics in convergence zones.

The continental Moho is not a first-order seismic discontinuity in many locations, as was traditionally thought. Rather, it appears to extend over a broad depth range in which the composition changes progressively from crustal metamorphic rocks to upper mantle peridotites. Reflectivity studies suggest that the Moho zone includes interlayered strips of both crustal and upper mantle rocks (Fuchs, 1986; Mooney and Brocher, 1987; Jarchow and Thompson, 1989). The influence of the Moho on mantle convection is profound. Among all of the interfacial regions within the Earth, only the core–mantle boundary supports a larger density increase. The density of the continental crust averages 2,600–2,700 kg m^{-3} while the density of the uppermost mantle is about 3,350 kg m^{-3}. Surface topography and Moho depth are well correlated at long wavelengths, indicating that continental topography is mirrored at the Moho. Buoyancy of continental rocks with respect to the mantle is the main reason why the continental crust resists subduction and remains on the Earth's surface for billions of years.

On the basis of tectonic history, continental regions are classified as shields, platforms, and mountain belts or orogens. Shields and platforms are also referred to as cratons. This division is reflected to some degree in the deep structure, but it is most evident in the surface geology. Shields are ancient, stable, slightly elevated regions, nearly free of sedimentation, and primarily composed of Precambrian igneous and metamorphic rocks that have remained
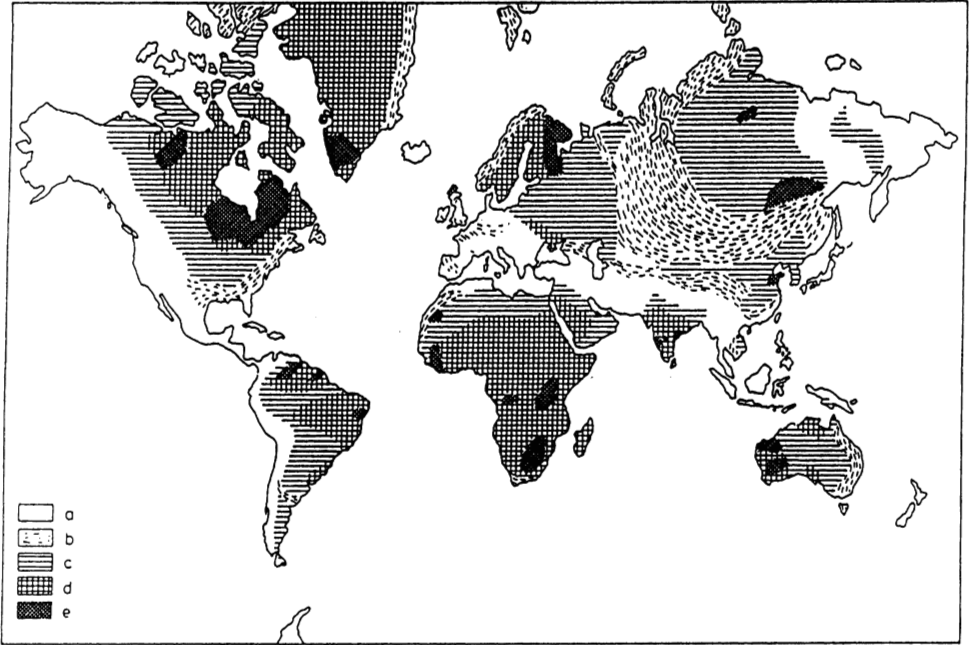
Figure 3.7. Age provinces of the continental crust. a = Mesozoic and Cenozoic tectonic belts, b = Paleozoic tectonic belts, c = Proterozoic platforms, d = Proterozoic shields, e = Archean shields. Adapted from figures in Miyashiro et al. (1982) and Meissner and Wever (1989).

undisturbed for 1 Gyr or longer. Platforms are areas covered by a relatively thin veneer of flat-lying sediments of Proterozoic and younger age, again with little or no deformation. The primary difference between shields and platforms is that platforms occasionally have subsided below sea level. Orogenic mountain belts with Phanerozoic age deformation tend to form curved sections surrounding shields and platforms, as shown in Figure 3.7. This concentric pattern of crustal ages is usually interpreted as evidence for continental growth by accretion.

A global model of the Earth's crust based on seismic refraction data published between 1948 and 1995 and a detailed compilation of ice and sediment thickness has been presented by Mooney et al. (1998). Crustal thicknesses from this model are shown in Figure 3.8, which prominently displays the difference in thickness of oceanic and continental crust. The thickest crust is found beneath the Tibetan Plateau (70 km) and the Andes. Anomalously thick crust in the oceans is associated with oceanic plateaus and hot spots.

3.4 The Upper Mantle

As discussed above, the existence of the Moho as the boundary between the crust and the mantle was inferred from seismic velocity observations early in this century. Until the 1960s there were two competing hypotheses for the nature of the Moho. The first was a compositional boundary between the silicic and basaltic rocks of the crust and the ultrabasic rocks of the mantle, principally peridotite, that have the appropriate seismic velocity for the
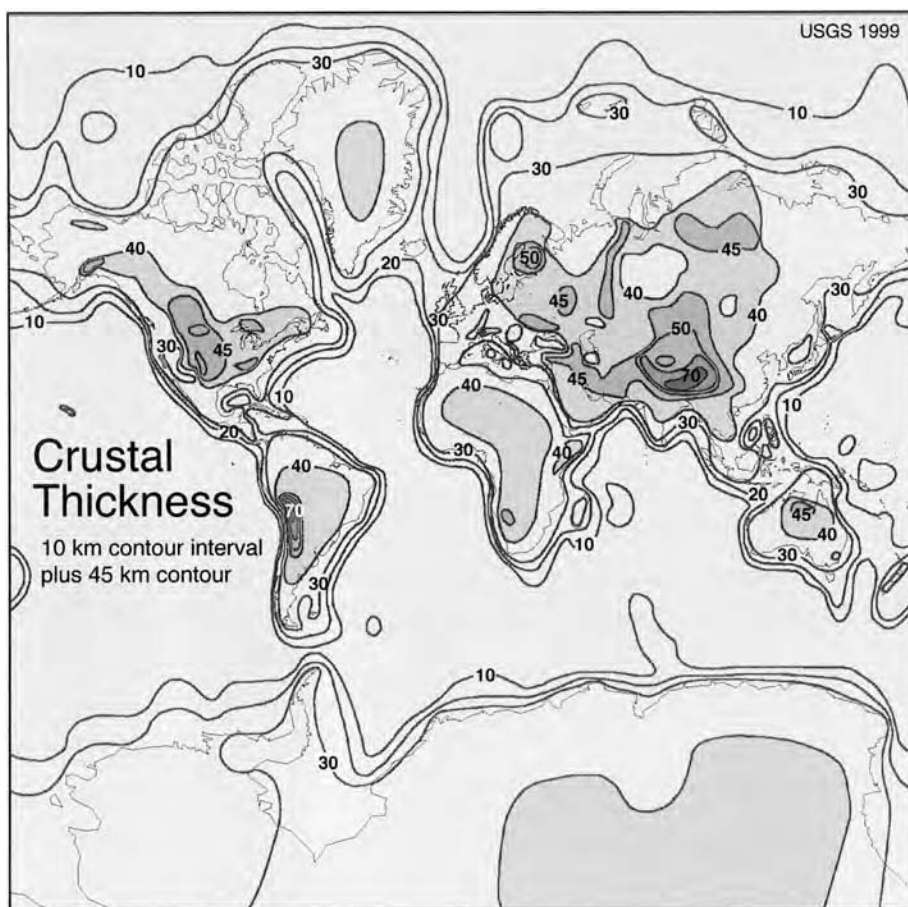
Figure 3.8. Global map of the thickness of the Earth's crust from the CRUST 5.1 model of Mooney et al. (1998). Crust thickness includes topography above sea level but not bathymetry.

For a color version of this figure, see plate section.

upper mantle. The second hypothesis was a phase boundary between basalt and its high-pressure equivalent eclogite. Like peridotite, eclogite has seismic velocities appropriate for the upper mantle.

With the general acceptance of plate tectonics, the compositional hypothesis for the Moho became favored. Confirmation came with the recognition of ophiolitic complexes as exposed sections of the oceanic crust. In Newfoundland, Cypress, Oman, and other places where the Moho is exposed at the Earth's surface, the basaltic and gabbroic rocks of the former oceanic crust are underlain by ultrabasic rocks with a mostly peridotitic composition.

3.4.1   Radial Structure of the Upper Mantle

Regional models of upper mantle structure derived from seismic surface waves commonly show very little increase in seismic velocities with depth for the first few tens of kilometers beneath the Moho. At greater depths, down to about 200 km, seismic velocities often decrease
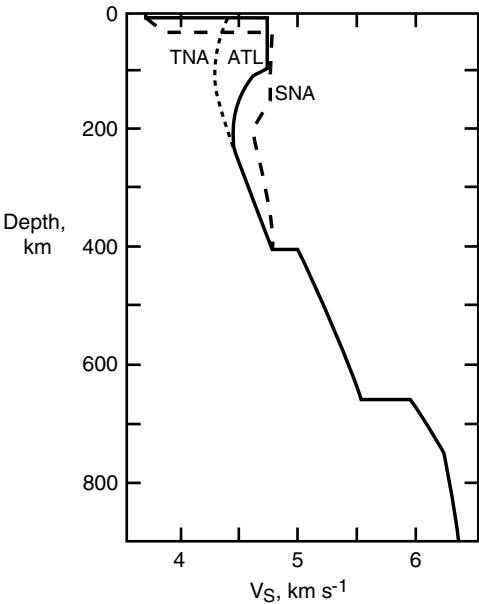
Figure 3.9. Regionalized upper mantle profiles of V_S from Grand and Helmberger (1984a, b). SNA = Canadian shield, TNA = tectonic western North America and East Pacific Rise, ATL = northeastern Atlantic.

with depth. We refer to the depth interval with nearly uniform velocity as the seismic lid (LID), and the depth interval with negative velocity gradients as the low-velocity zone (LVZ). This combination of structures is particularly evident in the oceanic upper mantle. Regional shear wave velocity profiles exhibiting these structures are shown in Figure 3.9.

The seismic LID roughly corresponds to the Earth's surface thermal boundary layer, the lithosphere. It has been speculated that the top of the low-velocity zone marks the transition in long-term rheology from the rigid lithosphere to the viscous asthenosphere. The upper part of the mantle LID has been depleted by the removal of the basaltic component to form the oceanic crust. The observed increase in seismic velocity with depth through the LID is attributed to this compositional change which dominates over the decrease in seismic velocity due to the temperature increase with depth through the LID. The mean density in the LID is known to be 3,300–3,400 kg m^{-3} from the isostatic balance between oceanic and continental lithosphere.

The LID and the LVZ are seismically anisotropic, particularly beneath the oceans, and this fact complicates the problem of determining their structure and depth extent. Seismic anisotropy in the upper mantle is determined by a variety of methods, including azimuthal variations of P$_n$ (P waves critically refracted along the Moho), shear wave birefringence (the dependence of wave speed on polarization direction), and comparisons between the dispersion of Rayleigh and Love surface waves.

The seismic anisotropy of the oceanic upper mantle was first measured by refraction studies (Morris et al., 1969; Raitt et al., 1969), which demonstrated that P$_n$ is about 5% faster in the direction of seafloor spreading than in the direction parallel to the ridge. This azimuthal seismic anisotropy is attributed to the statistical alignment of olivine crystals, with the fast *a*-axes parallel to the spreading direction and the *b*- and *c*-axes concentrated in a vertical plane parallel to the spreading axis (Christensen and Salisbury, 1979). A preferential grain alignment is thought to result from the strain field at spreading centers (see Chapter 4).

An implication of this hypothesis is that anisotropy is "frozen" into the oceanic LID at the time of its formation.

Seismic anisotropy is also present in the upper mantle beneath the continents, as revealed by shear wave splitting, a phenomenon analogous to optical birefringence. When a seismic shear wave propagates through anisotropic regions of the crust and upper mantle, it splits into two waves with orthogonal polarizations and slightly different propagation speeds. Observations of shear wave splitting have demonstrated that anisotropy in the upper mantle beneath tectonic regions is often consistent with the regional pattern of crustal deformation, an indication that the upper mantle is coupled to tectonic events on the continents (Silver, 1996).

On even larger scales, seismologists observe a systematic discrepancy between the dispersion of Love and Rayleigh surface waves, which can be explained if horizontally polarized shear waves propagate faster than vertically polarized shear waves in the upper 200 km of the mantle. The global average of this difference is small, about 2% (Forsyth, 1975; Dziewonski and Anderson, 1981), but it is nonetheless significant. It implies that the isotropy of the upper mantle is, on average, transverse. This form of seismic anisotropy is defined by a single axis of symmetry, with isotropic behavior in planes perpendicular to the axis, and is described by five independent elastic parameters. In the upper mantle, the symmetry axis is oriented in the vertical (radial) direction, so the isotropic surfaces are the horizontal (spherical) surfaces. Transverse isotropy can be produced by crystals oriented randomly on spherical surfaces, but aligned in the radial direction, or alternatively, by a structure consisting of fine-scale horizontal laminations.

Azimuthal anisotropy in Rayleigh waves provides a means for mapping the direction of anisotropy in the LID and LVZ worldwide. Whereas the LID seismic anisotropy records the fossil spreading direction, the direction of seismic anisotropy in the LVZ is probably not frozen, but instead reflects grain alignment in the present field of flow. In principle then, LVZ directional seismic anisotropy provides a tool for determining the direction of shear beneath plates. We shall return to this topic in our discussion of seismic tomography later in this chapter.

The importance of seismic anisotropy to the problem of determining vertical structure is due to the tradeoff between anisotropy and LID thickness in the inversion of surface wave data. Figure 3.10 shows the variation in oceanic LID thickness as a function of crustal age, as determined from surface wave dispersion with and without anisotropy. Isotropic models yield apparently thick LIDs (Forsyth, 1975), whereas anisotropic models give LID thicknesses only about one-half as large (Regan and Anderson, 1984; Kawasaki, 1986). The isotropic LID is very close to the predicted thermal boundary layer thickness (see Chapter 4), whereas the anisotropic LID is closer to the mechanical thickness determined from lithosphere flexure studies (Levitt and Sandwell, 1995) (see Chapter 5).

The LVZ is present as a global feature for S waves, although its most important characteristic is its regional variability, as shown in Figure 3.9. Shield regions have high shear velocity LIDs to a depth of about 300 km and little or no LVZ. Tectonically active areas on continents, on the other hand, may not have a well-defined LID, and instead have low velocities extending to about 200 km depth. The most systematic variations in LVZ structure are found beneath ocean basins, where the LID thickness increases approximately as the square root of crustal age (Figure 3.10), similar to the increase in thermal boundary layer and elastic flexure thicknesses. This trend indicates a direct connection between seismic and thermal structure in the oceanic upper mantle. In the oceanic LID, temperature, pressure, and compositional gradients combine to produce a nearly constant seismic velocity depth
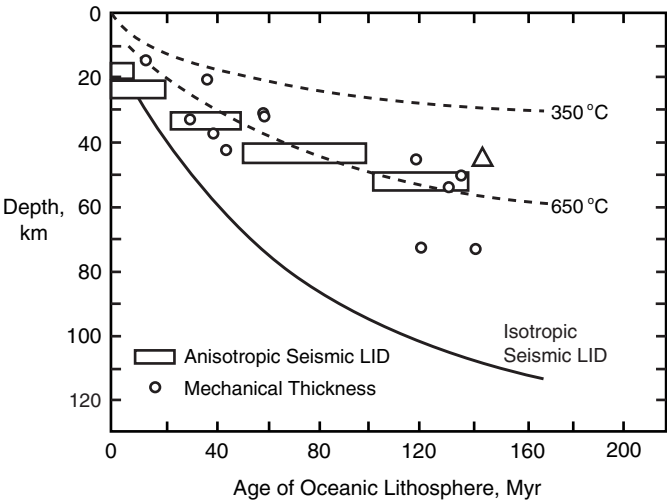
Figure 3.10. Oceanic LID thicknesses and mechanical lithosphere thicknesses versus crustal age (modified from Regan and Anderson, 1984). Boxes are LID thicknesses obtained by Regan and Anderson (1984) from surface wave dispersion assuming transverse anisotropy and anelastic dispersion. The solid curve represents seismic LID thickness estimates assuming isotropy. Circles are the thicknesses of the mechanical lithosphere computed by Levitt and Sandwell (1995) from elastic flexure analysis of lithospheric bending at subduction zones. Mechanical thicknesses represent thicknesses of the part of the lithosphere with long-term strength to resist deformation and are obtained from elastic thicknesses using a yield strength envelope model. The isotherms are from the thermal boundary layer model for the oceanic lithosphere given in Chapter 4.

profile; the thickness of the LID increases with increasing crustal age due to the conductive cooling of the lithosphere.

The LVZ beneath ocean basins is often interpreted as the intersection of the geotherm with the solidus of mantle rocks, and hence it marks the depth of partial melting. Partial melting in the LVZ is consistent with the distribution of seismic attenuation in the upper mantle. The quality factor Q (inverse attenuation rate) has a sharp minimum in the LVZ, particularly for S waves, as shown in Figure 3.11. Partial melts, even in low concentrations, strongly increase attenuation (decrease Q), particularly for seismic shear waves. The melt phase is concentrated at grain boundaries and tends to form an interconnected fluid network, which can drastically increase the rate of energy dissipation in seismic shear waves.

***Question 3.1:** Is the LVZ partially molten?*

We have already noted that seismic shear wave velocities are much larger beneath continental cratonic regions (shields and platforms) than beneath tectonically active areas. Cratons evidently have "roots" extending into the upper mantle. There are a number of important unsolved problems concerning these roots, including their depth extent, possible compositional and rheological differences with respect to the rest of the upper mantle, and the implication that they are tightly bound to the overlying crust. Jordan (1975, 1981, 1988) introduced the term tectosphere to distinguish these roots from the rest of the lithosphere,

Figure 3.11. Profiles of shear wave velocity V_S and shear wave attenuation factor Q_S versus depth, showing oceanic LID–LVZ structure (after Anderson and Given, 1982).
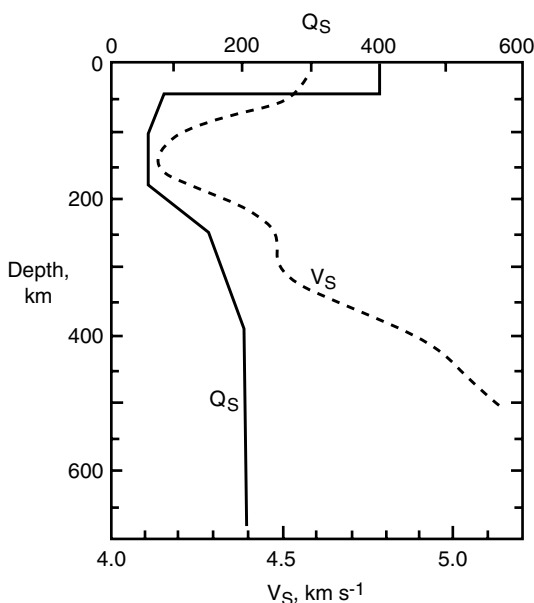
and has argued that they must be cold and basalt depleted in comparison with surrounding mantle. Low temperatures provide the mechanical strength to bind the root to the overlying crust, and basalt depletion results in a residual harzburgite composition with a lower density than the less depleted mantle. Since isostatic compensation of the continental crust is nearly complete at the Moho, the tectosphere must have the same average density as less depleted mantle, and this requires cancellation of the effects of temperature and basalt depletion on density in the tectosphere. While there are strong indications of mantle roots beneath cratons, their depth extent remains controversial. Sacks and Snoke (1984) have reviewed the evidence from seismology, in an attempt to detect patterns in continental root thickness. They find root thicknesses in the range 150–250 km, certainly greater than the oceanic LID thickness, but not as great as the 400–600 km originally proposed by Jordan (1975). Global seismic tomographic images show continental roots extending 300 km or more into the upper mantle, but the maximum depth of these structures is still poorly resolved (e.g., Su et al., 1994).

Question 3.2: *What is the thickness of the continental lithosphere? Do the continents have roots that extend to depths of several hundred kilometers in the mantle?*

One problem in detecting deep continental roots is that they may be masked by the Lehmann seismic discontinuity at about 220 km depth. The Lehmann seismic discontinuity is given in Earth model PREM (Figure 3.3) as an approximately 2% increase in velocity. It is often missing in profiles derived from body waves (Shearer, 1990), suggesting that it may be absent in some places, particularly beneath ocean basins. Gaherty and Jordan (1995) have

proposed that the Lehmann discontinuity represents the base of the anisotropic portion of the continental root, which would account for its absence in the oceanic upper mantle. The disappearance of anisotropy with depth could be a rheological effect (Karato, 1992; Gaherty and Jordan, 1995).

Question 3.3: *What is the nature of the Lehmann seismic discontinuity at a depth of 220 km?*

3.4.2   Upper Mantle Composition

The composition of the upper mantle is constrained by density, seismic velocities, and seismic anisotropy. It is also constrained by compositions found in ophiolite complexes and mantle xenoliths carried to the surface in kimberlitic and alkali basaltic eruptions. Furthermore, the compositions of the crust, mantle, and core must be consistent with cosmological estimates and, in addition, it is usually assumed that the upper mantle composition produces MORB upon 10–20% partial melting.

First, it is necessary that any model for mantle composition satisfy cosmochemical constraints. It is generally accepted that the meteorites classified as CI chondrites represent primordial building blocks of the Earth. Their mean bulk composition, with the exception of the volatile elements, is generally consistent with a mantle that is primarily peridotite and a core that is primarily iron, as shown in Table 3.3 (McDonough and Sun, 1995). Table 3.4 compares the average composition of CI chondrites with the so-called pyrolite composition, a peridotite model of the mantle introduced by Ringwood (1975) to explain the seismic, petrologic, and mineralogic properties of the upper mantle. Except for small differences in silicon and magnesium, the two compositions are very similar.

Relatively few silicate and oxide compounds are structurally stable at mantle pressures and temperatures, and this restricts the bulk composition to a mixture of olivine, pyroxenes, garnet, magnesiowüstite and the quartz polymorph stishovite. Chemical formulas and estimated

Table 3.3. *Bulk Earth Composition[a]*

Element	Mantle wt (%)	Core wt (%)	Bulk Earth wt (%)
Al	2.35		1.59
Ca	2.53		1.71
Mg	22.8		15.4
Si	21.0		14.2
Fe	6.26	87.5	32.7
Cr	0.263	0.95	0.49
Ni	0.196	5.40	1.89
Mn	0.105	0.50	0.24
Na (ppm)	2,670		1,800
V (ppm)	82	120	95
Co	0.0105	0.26	0.0915
P	0.009	0.050	0.17

[a] McDonough and Sun (1995).

Table 3.4. *Chondrite Versus Pyrolite Composition*[a]

Component	CI Chondrites wt (%)	Pyrolite wt (%)
SiO_2	49.9	45.0
TiO_2	0.16	0.201
Al_2O_3	3.65	4.45
Cr_2O_3	0.44	0.384
MnO	0.13	0.135
FeO	8.0	8.05
NiO	0.25	0.25
MgO	35.15	37.8
CaO	2.90	3.55
Na_2O	0.34	0.36
K_2O	0.022	0.029

[a] McDonough and Sun (1995).

Table 3.5. *Pyrolite Model for Upper Mantle Composition*[a]

Minerals	Compositions	Pyrolite wt (%)
Olivine	$(Mg, Fe)_2SiO_4$	57.9
Orthopyroxene	$(Mg, Fe)SiO_3$	13.5
Clinopyroxene	$([Ca, Mg]_2, NaAl)Si_2O_6$	16.3
Garnet	$(Mg, Fe, Ca)_3Al_2Si_3O_{12}$	12.3

[a] Ringwood (1975); Green and Falloon (1998).

mantle abundances of these minerals are listed in Table 3.5. Olivine is a complete solid solution of Mg and Fe silicates with end members fayalite, Fe_2SiO_4, and forsterite, Mg_2SiO_4. Mantle pyroxenes are orthopyroxene and clinopyroxene. Orthopyroxene is the limited solid solution, $(Mg, Fe)SiO_3$. The magnesium end member is enstatite; the iron end member is unstable. Clinopyroxene is a pyroxene solid solution including Ca and Al. End members are diopside, $MgCaSi_2O_6$, hedenbergite, $FeCaSi_2O_6$, and jadite, $NaAlSi_2O_6$. The chemical formula for garnet is $(Mg, Fe, Ca)_3Al_2Si_3O_{12}$. Among all possible rock assemblages composed of these minerals, only peridotites (olivine + pyroxene) and eclogites (pyroxene + garnet) are commonly found in mantle-derived samples. Eclogite is isochemical with basalt, and transforms to basalt at depths of less than about 80 km within the Earth.

The problem of determining the upper mantle composition from the composition of mantle xenoliths, plus the geophysical constraints, has been thoroughly treated (Boyd and Meyer, 1979; Ringwood, 1979; Anderson, 1989), although there is some disagreement on the most likely composition. Xenoliths from diamond-bearing kimberlite pipes, with source depths of 150 km or greater, are mostly peridotites; only a small fraction are eclogites. The most common peridotitic xenoliths are garnet lherzolites (about 60% olivine, the remainder orthopyroxene + clinopyroxene + garnet) and harzburgites (olivine + orthopyroxene), with eclogites and spinel lherzolites (spinel in place of garnet) being widespread but less abundant. If one assumes that the sampling by kimberlites is unbiased, then the upper mantle is composed primarily of peridotite, with eclogite an important but less abundant constituent.

This inference is consistent with seismic velocities and densities, and the abundance of olivine provides an explanation for seismic anisotropy in the upper 200 km of the mantle. In contrast, eclogite has a density of 3,500 kg m^{-3} and is only weakly seismically anisotropic. It is therefore unlikely that eclogite is a dominant upper mantle assemblage.

For the purpose of determining the physical properties of the upper mantle, garnet peridotite is probably a good first-order approximation to the bulk composition. However, it is not completely satisfactory as a chemical model, because upper mantle rocks must be able to produce MORB upon partial melting. It is necessary to include some eclogite in the upper mantle composition in order that it serve as a source for basalt. A mantle rock that can produce MORB by 10–20% partial melting is said to be fertile. Mantle rocks from which the basaltic component has been extracted are said to be infertile. Most peridotites found in xenoliths are too refractory to produce large amounts of basaltic magma, i.e., they are too depleted in the pyroxenes due to prior partial melting (Hawkesworth et al., 1990).

There are basically two classes of models for upper mantle composition that can satisfy the geophysical constraints and also explain the origin of MORB. One class of models is based on the assumption that the dominant upper mantle rock type is a mixture of peridotite and eclogite. Ringwood (1975) proposed that its bulk composition corresponds to about three parts refractory peridotite plus one part eclogite, a hypothetical assemblage he termed pyrolite. The upper mantle abundances of minerals according to the pyrolite hypothesis are given in Table 3.5 (Ringwood, 1975; Green and Falloon, 1998). Partial melting of pyrolite beneath oceanic spreading centers leads to a compositionally stratified oceanic lithosphere, with basaltic ocean crust overlying refractory peridotites, which grade downward into pyrolite.

A second class of models for upper mantle composition is based on the assumption that the upper mantle is stratified. Dickinson and Luth (1971) proposed that the subducted oceanic crust transformed to eclogite and gravitationally separated to form the lower mantle beneath 700 km depth. A primary problem with this model is that the upper mantle would be so depleted that the oceanic crust could not be produced from it. A variant on this hypothesis has been proposed by Anderson and Bass (1984). They suggest that the entire mantle above the 410 km seismic discontinuity has the composition of the kimberlite xenoliths – mostly refractory peridotites – and that the source layer for MORB lies deeper in an eclogite-rich layer within the transition zone. In their model, the bulk composition of the upper mantle is too depleted to produce voluminous basaltic magmas and the transition zone is assumed to be the MORB source region. However, there are serious problems with any stratified mantle model. In order to maintain the stratification each layer would have to convect essentially independently. Significant transport of material across a stratification boundary would rapidly mix and homogenize the mantle. Also, it is difficult to envision a stratified mantle that could generate a basaltic oceanic crust of near-uniform composition. Such a crust is a natural consequence of ascending convection in an upper mantle that is a mixture of peridotite and eclogite.

Figure 3.12 is a comparison of several regionalized seismic velocity depth profiles with velocities calculated by Anderson and Bass (1984) along adiabats for the pyrolite model of upper mantle composition. Compressional V_P and shear V_S wave velocities are given for continental shields, the northwest Atlantic, the western Pacific, and tectonic rises (northern part of the East Pacific Rise, Gulf of California, and southern California). The expected seismic velocities for a pyrolite composition mantle at various potential (zero pressure) temperatures are plotted. The adiabatic pressure dependence of these velocities is obtained without taking into account partial melting. The portions of these curves beneath the solidus curves are in the partial melt field. The high velocities in the seismic lid and the low velocities
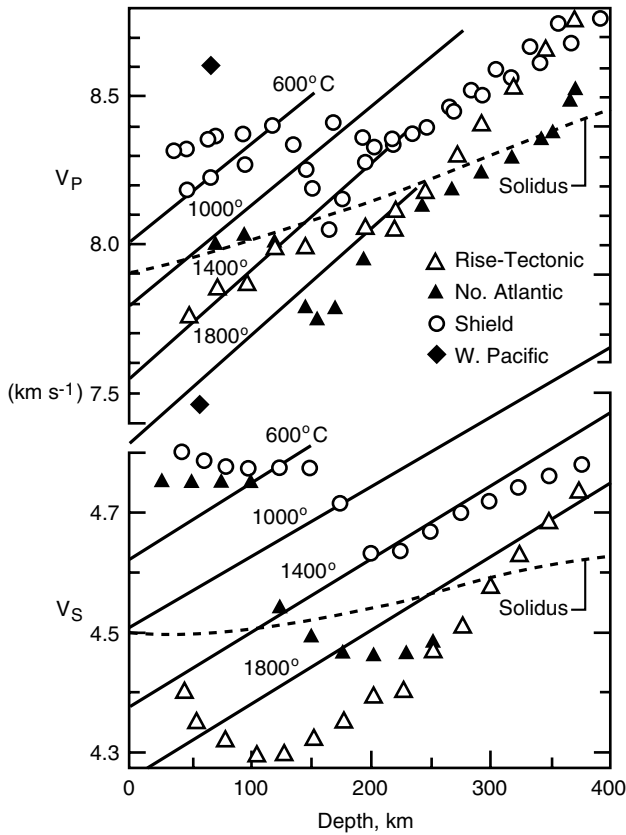
Figure 3.12. Compressional and shear wave velocities (V_P, V_S) are given as a function of depth for several regions (Anderson and Bass, 1984). The solid lines are the expected seismic velocities for a pyrolite composition mantle at various potential (zero pressure) temperatures. The adiabatic pressure dependence of these calculated velocities does not take into account partial melting. The portions of the data and curves beneath the dashed solidus lines lie in the partial melt field. The symbols are regionalized upper mantle seismic velocity profiles.

in the asthenosphere (LVZ) are clearly illustrated. The increase in velocity, from the LVZ down to 400 km depth, is reasonably consistent with a homogeneous composition under isentropic compression.

Another point revealed by the comparison in Figure 3.12 is the likelihood that the LVZ is caused by partial melting. In general, steep thermal gradients reduce V_P more than V_S, while small amounts of intergranular melts tend to reduce V_S more than V_P (Anderson, 1989). As shown in Figure 3.12, the LVZ is more pronounced for V_S, which is consistent with partial melting being the cause of the velocity decrease, rather than steep thermal gradients. Furthermore, the temperature distributions implied by the data in Figure 3.12 seem to require the presence of partial melts. Adiabats fitted to the seismic velocity profiles for spreading centers and tectonically active areas in Figure 3.12 lie above the solidus in the interval 50–300 km depth. It is worth emphasizing that this interpretation suggests that mid-ocean ridges and tectonically active regions such as western North America are underlain by a partially molten layer. Regardless of whether it is due to steep thermal boundary layer temperature gradients, or partial melts, or both, the LVZ is a product of mantle convection.

3.5 The Transition Zone

3.5.1 The 410 km Seismic Discontinuity

In 1926 Byerly (1926) discovered a decrease in slope of the P-wave travel time curve at angular distances near 20°. It implied a sharp jump in V_P, which became known as the "20° seismic discontinuity." Bullen's early Earth models included this seismic discontinuity, and Bernal (1936) suggested it marked a solid-state phase transformation from the olivine to the denser spinel structure. Francis Birch took up the problem of interpreting mantle seismic structure in terms of composition; his efforts culminated in a landmark 1952 paper in which he applied an equation of state, derived from finite-strain theory, to Bullen's Earth model B. Birch (1952) concluded that an upper mantle composition of olivine, pyroxene, and garnet is consistent with the seismic data, and proposed that the lower mantle is composed of a mixture of dense oxides such as periclase (MgO). He also concluded that the transition zone represents a succession of pressure-induced polymorphic phase changes. At the time there was neither the seismic nor the mineral physics data to fully test Birch's proposals. In the 1960s, seismic array studies revealed that the transition zone has a fine structure including two sharp discontinuities near depths of 410 and 660 km (Johnson, 1967). At nearly the same time, the olivine–spinel phase transformation was observed in laboratory experiments at approximately the correct pressure to explain the 410 km seismic discontinuity (Akimoto and Fujisawa, 1968; Ringwood and Major, 1970). The presently accepted properties of this phase change include a $0.5 \, \text{km s}^{-1}$ jump in V_P and a 0.2–$0.3 \, \text{km s}^{-1}$ jump in V_S, centered near a depth of 410 km and distributed over a depth interval of 15 km or less (Fukao, 1977; Leven, 1985; Shearer, 1990, 1991). The lateral variation in depth is rather small, generally within ± 10 km of the global average depth (Shearer, 1991). The accompanying density jump is in the range 200–$300 \, \text{kg m}^{-3}$, as shown in Figure 3.3.

Phase relations in the transition zone and in the lower mantle are primarily controlled by two mineralogies, olivine and pyroxene–garnet. Garnet peridotite, the dominant upper mantle assemblage, consists of these in roughly a $3:1$ proportion, implying that the phase transformation of olivine should control the character of the 410 km seismic discontinuity.

The solid solution in olivine is usually expressed in terms of the magnesium number X_{Mg}, the molar concentration of Mg normalized by the molar concentration of Mg plus Fe, expressed as a percent. The upper mantle density $3{,}350 \, \text{kg m}^{-3}$ is matched by olivine with a composition $(\text{Mg}_{0.89}, \text{Fe}_{0.11})_2\text{SiO}_4$, or $X_{\text{Mg}} \simeq 89$. Early experiments on the iron end member fayalite verified the transformation from the low-pressure α phase, with orthorhombic structure, to the cubic structure γ phase at about 7 GPa. The phase diagram for olivine is shown in Figure 3.13. At transition zone temperatures of $1{,}600$–$1{,}900$ K, $X_{\text{Mg}} = 89$, olivine transforms to an intermediate β-modified spinel phase at 13–14 GPa, and from the β-modified spinel phase to the γ-spinel phase at 17–18 GPa. The transformation α olivine $\rightarrow \beta$-modified spinel occurs near 13.5 GPa, 405 km depth, at about 1,700 K (Akaogi et al., 1989; Katsura and Ito, 1989), and is the most likely candidate to explain the 410 km seismic discontinuity.

In addition to depth, the width of the mixed-phase region, where α olivine and β-modified spinel coexist, is crucial in identifying it with the 410 km seismic discontinuity. Experimental evidence indicates that the α olivine $\rightarrow \beta$-modified spinel transformation occurs over a narrow pressure interval. For olivine with magnesium number 89, Katsura and Ito's (1989) measurements indicate a 10–20 km thickness for the α olivine $\rightarrow \beta$-modified spinel phase change in the temperature range $1{,}600$–$1{,}900$ K, with the width of the two-phase loop
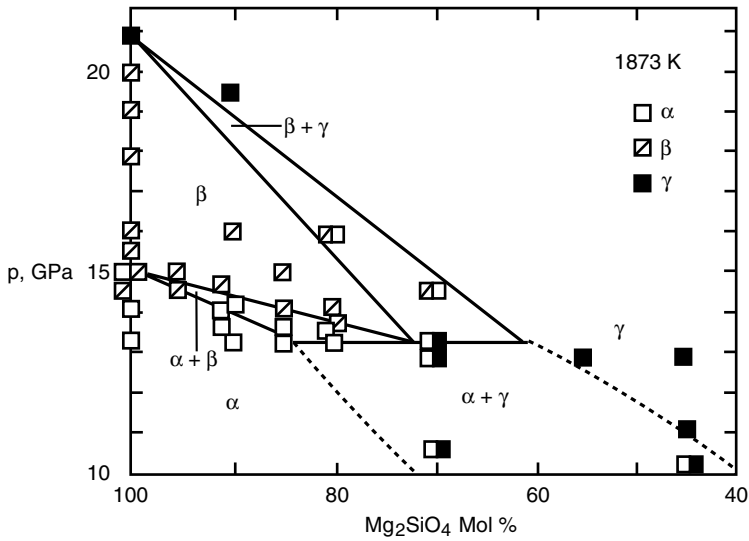
Figure 3.13. Phase diagram for the system $(Mg, Fe)_2SiO_4$ at pressures of 10–21 GPa and $T = 1,873$ K (Katsura and Ito, 1989).

decreasing with increasing temperature. This is slightly thicker than most estimates of the depth interval of the 410 km seismic discontinuity, but within the present uncertainties. The experimentally determined Clapeyron slopes are positive, indicating that this reaction is exothermic. Estimates of the Clapeyron slope vary from $dp/dT \simeq 2.5\,\mathrm{MPa\,K^{-1}}$ (Katsura and Ito, 1989) to $dp/dT \simeq 1.5\,\mathrm{MPa\,K^{-1}}$ (Akaogi et al., 1989).

The P-wave velocity jump accompanying α olivine $\to$ β-modified spinel has been estimated at 6.6% by Bina and Wood (1987); it is possible to account for the observed seismic velocity increase if the upper mantle is about 75% olivine. Thus, the experimental evidence supports the interpretation of the 410 km seismic discontinuity as the α olivine $\to$ α olivine $+\beta$-modified spinel $\to$ β-modified spinel phase transformation. The topography of the 410 km seismic discontinuity has been mapped by Flanagan and Shearer (1998a, b, 1999) as discussed further below.

The degree to which the olivine composition accounts for seismic velocities through the transition zone is shown in Figure 3.14, which compares equation of state calculations by Bina and Wood (1987) with PREM seismic velocities. The variation in seismic velocity can be well represented by a uniform composition from a depth of 410 km to the depth of the 660 km seismic discontinuity. A similar comparison has been made by Weidner (1985) using a pyrolite model composition, with the same result: there is no compelling reason to invoke compositional variation to explain the transition zone structure above a depth of 660 km.

However, this does not mean that compositional variations are necessarily absent in the transition zone. The seismic properties of other mantle compounds are similar to those of olivine at transition zone pressures. This is the case for pyroxenes, the second most abundant mineral in the upper mantle. There are several pressure-induced phase transformations in pyroxenes that are relevant to transition zone structure (Akaogi et al., 1987). At high temperatures, enstatite transforms to majorite, an Al-deficient garnet structure; at low temperatures, the transformation is from enstatite to β spinel plus stishovite (the high-pressure quartz polymorph, SiO_2 with Si in sixfold coordination), then to γ spinel plus stishovite.
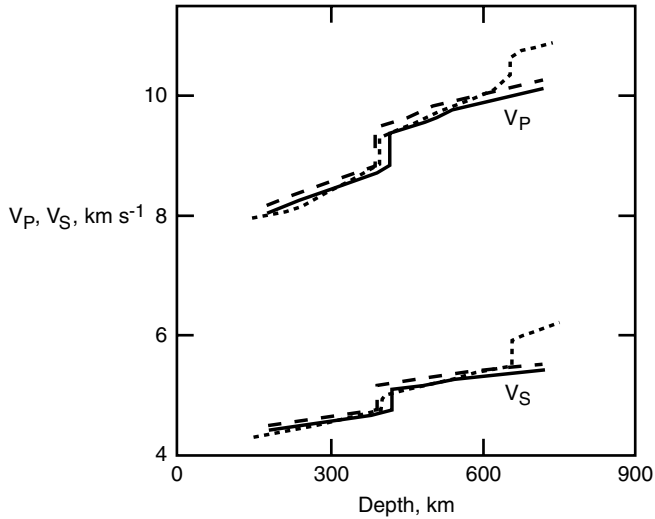
Figure 3.14. Depth profiles of V_P and V_S in the transition zone given by Bina and Wood (1987) for $X_{Mg} = 90$ olivine composition along 1,700 K (dashed) and 2,000 K (solid) isotherms. Dotted curves are the regionalized velocity profiles for North America given by Grand and Helmberger (1984a, b).

The approximate stability fields are shown in Figure 3.15. Elastic velocities are not as well determined for the pyroxene system as for the olivine system, but it is likely that pyroxenes would not appreciably change the seismic velocity profiles in Figure 3.14 if they were present in levels less than 30%. Thus, small variations in the relative abundance of olivine and pyroxene through the transition zone cannot be excluded.

Phase transformations in the pyroxene system are not expected to be as sharp as those in the olivine system, because pyroxene compounds accommodate greater amounts of "minor abundant" cations, such as Ca and Al, and these tend to spread out phase changes over a wider pressure interval. By virtue of their sharpness, transformations in olivine compounds can produce seismic discontinuities in the mantle, whereas transformations in pyroxenes are too gradual. Thus, the phase changes in olivine may be responsible for the 410 km and 660 km seismic discontinuities, while phase changes in pyroxene may account for the steep velocity gradients elsewhere in the transition zone (Figure 3.3).

3.5.2 The 660 km Seismic Discontinuity

The seismic discontinuity near 660 km depth contains the sharpest changes in seismic properties within the mantle. Over an interval of about 5 km or perhaps even less, V_P and V_S increase by 0.4–0.6 and 0.3–0.4 km s^{-1}, respectively, and the density ρ increases by about 400 kg m^{-3}, or about 9% according to Earth model PREM.

The sharpness of this interface was originally inferred from observations of underside reflections, precursors to the seismic phase P′P′ (PKPPKP). Richards (1972) and Benz and Vidale (1993) have used P′P′ precursor amplitudes to infer a thickness of 4 km or less for the seismic discontinuity. Lees et al. (1983) argued that such reflections would not occur if the thickness of the seismic discontinuity were much greater than a few kilometers. The interpretation of P′P′ precursors as underside reflections from the 660 km seismic discontinuity

Figure 3.15. Phase stability fields in temperature–pressure coordinates for $MgSiO_3$. After Sawamoto (1987).
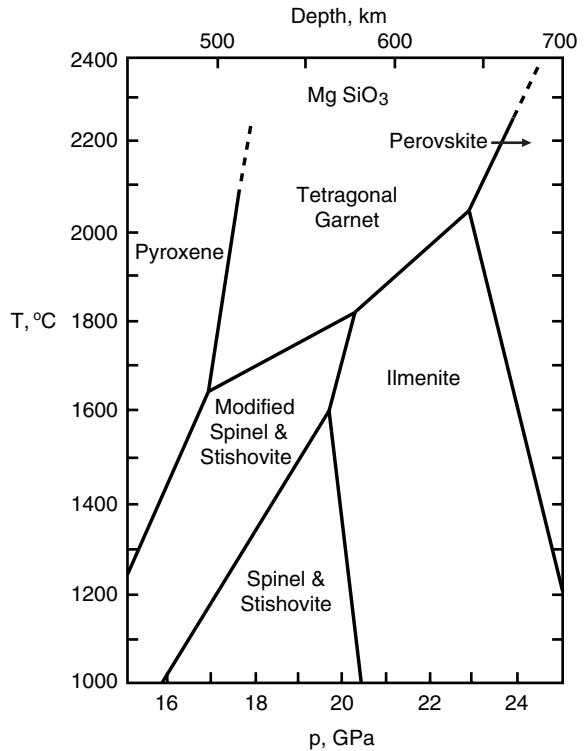
has been questioned (Estabrook and Kind, 1996), but the conclusion that this discontinuity occurs over a very narrow pressure range persists.

In the literature, this seismic discontinuity is variously located between depths of 650 km and 670 km. The discrepancy is partly an indication of the difficulty in determining its absolute depth precisely, and partly due to lateral variations in its absolute depth (Shearer and Masters, 1992). In this text, we shall refer to it simply as the "660 km seismic discontinuity."

Question 3.4: *What is the nature of the 660 km seismic discontinuity?*

The composition of the transition zone in general, and the nature of the 660 km seismic discontinuity in particular, represents one of the most important constraints on any model of mantle convection. No other part of the structure of the mantle has been subject to as much interpretation and conjecture. The seismicity of the Earth terminates abruptly near this depth, a fact that many geophysicists interpreted to mean that subducted lithospheric slabs are unable to penetrate across the discontinuity, and thus remain in the upper mantle. This interpretation led to the idea of a two-layer mantle, in which the 660 km seismic discontinuity divides the mantle into two convecting regions, with only limited mass transfer occurring across the discontinuity.

What property of the 660 km seismic discontinuity would make it possible to stop subducted slabs from entering the lower mantle? Models of mantle convection described in Chapter 9 indicate that the endothermic pressure-induced phase transformations near 660 km

depth inhibit slab penetration, but do not provide a permanent barrier. In addition, the inferred increase in viscosity near this depth discussed in Chapter 5 may affect the structure of the flow by causing slabs to decelerate and buckle, but would not prevent them from sinking into the lower mantle. Only an intrinsic density increase, due to changes in bulk composition, can provide a sufficiently strong restoring force to keep negatively buoyant lithosphere in the upper mantle. Thus, the dynamical significance of the 660 km seismic discontinuity revolves around the question of whether it is simply a phase transformation, or whether there is also a change in bulk composition.

Question 3.5: *Is the composition of the lower mantle different from that of the upper mantle?*

There are basically two methods to approach the problem of determining lower mantle composition, and both require making a priori assumptions about upper mantle composition. One method involves experimental determinations of the phase equilibrium diagram for the important mantle assemblages over the pressure and temperature range of the 660 km seismic discontinuity, leading to the determination of which composition has the phase transformation properties that best explain the observed depth and structure of the seismic discontinuity. The other method is to subject assumed upper mantle compositions to the pressures and temperatures of the lower mantle, and determine if any such compositions are consistent with lower mantle seismic properties over the wide pressure range of the lower mantle. Both approaches have been tried, but only recently have there been enough mineral property data to make them meaningful. Previously, models of lower mantle mineralogy were heavily influenced by Birch's "mixed oxide" proposal, in which stishovite SiO_2 and magnesiowüstite (Mg, Fe)O were presumed to be the dominant phases. It was found that the lower mantle is a few percent denser than an oxide mixture that is isochemical with pyrolite (Ringwood, 1970; Davies and Dziewonski, 1975). This difference was interpreted as being due to iron enrichment in the lower mantle relative to the upper mantle (Anderson, 1970).

The situation changed with the discovery by Liu (1974) of a high-density silicate perovskite structure for $MgSiO_3$. Since then, it has been shown that nearly all mantle silicates transform to the perovskite structure at lower transition zone pressures. Since the density of $MgSiO_3$-perovskite is 3–5% higher than the isochemical oxide mixture, the presence of silicate perovskites below the 660 km seismic discontinuity removes the requirement for a large increase in lower mantle iron content in order to reconcile upper mantle composition with lower mantle seismic properties. However, there is still room for a small difference between the iron and silica contents of the upper and lower mantle.

Experimental work on the high-pressure phase diagram of mantle silicates and their transformation to perovskite structures involves diamond anvil pressure devices and, particularly in Japan, large-volume split-anvil pressure devices. Figure 3.16 shows phase diagrams for the olivine system in the pressure range 22–26 GPa, at 1,373 K and 1,873 K, as determined by Ito and Takahashi (1989). γ spinel with an upper mantle composition $X_{Mg} \simeq 89$ transforms to a mixture of magnesium perovskite and magnesiowüstite over a very small pressure interval, less than 0.15 GPa, which corresponds to 4 km or less in depth. Unlike the exothermic olivine–spinel reaction, this reaction is endothermic and has a negative Clapeyron slope, $dp/dT \simeq -2.5 \, \text{MPa K}^{-1}$ (Ito and Takahashi, 1989; Bina and Helffrich, 1994). The reaction
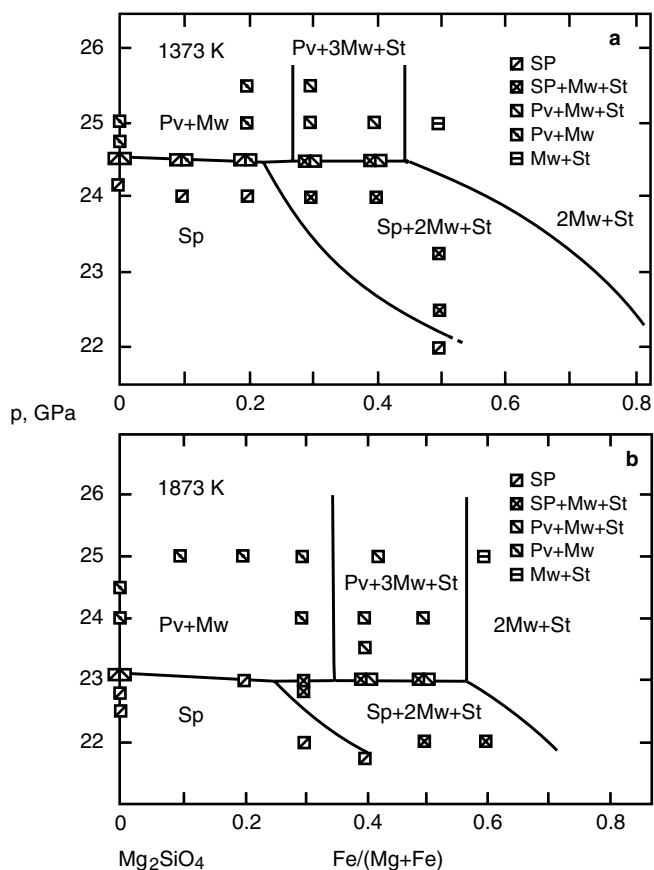
Figure 3.16. Phase diagram for the system $(Mg, Fe)_2 SiO_4$ between 22 and 26 GPa at (a) 1,373 K and (b) 1,873 K, determined by Ito and Takahashi (1989). Sp = spinel, Pv = perovskite structures, Mw = magnesiowüstite, St = stishovite.

pressure is nearly independent of iron content for $100 > X_{Mg} > 80$, the full range of possible mantle compositions. At 1,873 K, a reasonable estimate for the average temperature of that region of the mantle, the phase change occurs at 23.1 GPa, almost precisely at a depth of 660 km. At 1,373 K, more representative of temperatures in a subducting slab, the reaction occurs at 24.5 GPa, approximately 40 km deeper. This variation is nearly equal to the seismically observed variation of depths to the seismic discontinuity. Shearer and Masters (1992) and Flanagan and Shearer (1998a) have mapped the large-scale topography of the discontinuity (Figure 3.17), and find nearly 30 km of relief with the greatest depressions located beneath subduction zones. Detailed seismic studies of the depth of the 660 km seismic discontinuity beneath the Tonga (Richards and Wicks, 1990; Niu and Kawakatsu, 1995) and Izu-Bonin (Vidale and Benz, 1992; Wicks and Richards, 1993; Castle and Creager, 1997) slabs indicate depressions of the discontinuity up to 50–60 km below its nominal depth. These results offer a compelling argument that the 660 km seismic discontinuity marks the transformation of γ spinel to magnesium perovskite and magnesiowüstite in the olivine system. They argue strongly against the association of a major chemical compositional change with the 660 km seismic discontinuity since a chemical discontinuity should be locally dynamically

Depth (km)
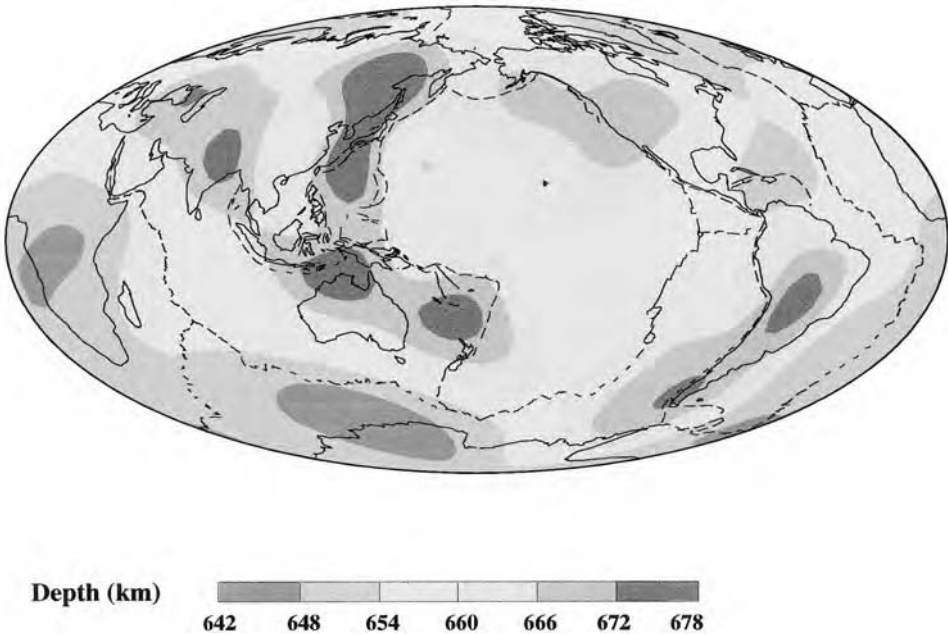
| 642 | 648 | 654 | 660 | 666 | 672 | 678 |

Figure 3.17. Depth of the 660 km seismic discontinuity. After Flanagan and Shearer (1998a).

For a color version of this figure, see plate section.

depressed by 100–300 km in subduction zones (Christensen and Yuen, 1984; Kincaid and Olson, 1987).

Perovskite-forming reactions also occur in the more silica-rich pyroxene system, but as is found for α olivine $\rightarrow$ β-modified spinel reactions, the presence of Ca and Al tends to spread these reactions over a wider pressure interval. The reaction of greatest interest is garnet dissociating to $MgSiO_3$ and $CaMgSiO_3$ perovskites plus Al_2O_3. This transformation begins at approximately the same pressure as in the olivine system, but it occurs over a broad pressure range, about 2–3 GPa (Irifune and Ringwood, 1987; Ito, 1989), and consequently it is spread out over a 50–100 km depth interval in the mantle below the 660 km seismic discontinuity. We conclude that the combination of perovskite-forming reactions in olivine and pyroxene systems can account for most of the fine structure in the lower portion of the transition zone contained in seismic Earth models such as PREM. The olivine reaction produces the reflecting discontinuity at 660 km depth, and the pyroxene reactions produce the region of high velocity gradients extending down to about 750 km depth, as shown in Figure 3.3.

While it is possible to explain the 660 km seismic discontinuity by perovskite-forming reactions, there is evidence for subtle compositional variations located somewhere within the transition zone. Over most of the pressure range of the lower mantle, from about 30 to 110 GPa, the Adams–Williamson condition (3.2.9) is satisfied to within 1% (Shankland and Brown, 1985) and the assumptions of homogeneity and adiabaticity are valid to a good approximation. It is therefore possible to determine if particular models of upper mantle composition, subjected to lower mantle pressures and temperatures, match the profiles of seismic velocity and density through the lower mantle. In order to make this comparison, it
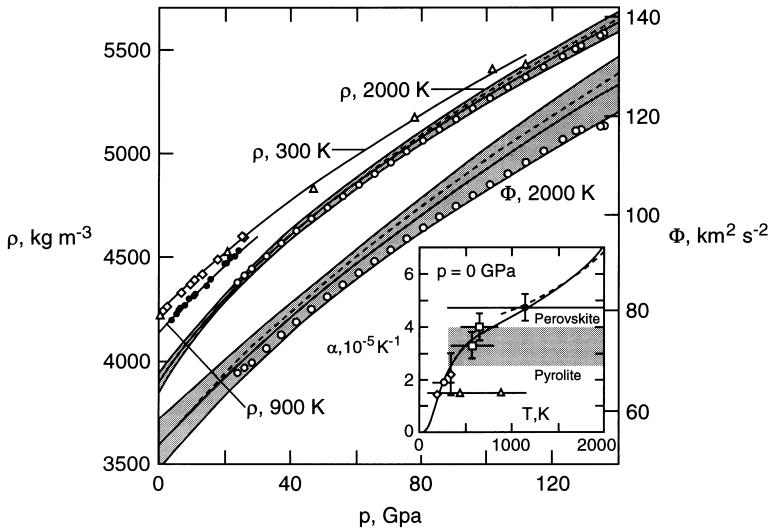
Figure 3.18. Comparison of density ρ and seismic parameter Φ in $(Mg_{0.9}Fe_{0.1})SiO_3$ perovskite with ρ and Φ in the lower mantle (modified from Stixrude et al., 1992). Perovskite density measurements at room temperature (300 K) are shown as open diamonds and triangles; measurements at 900 K are shown as solid circles. Solid and dashed lines represent extrapolations using equations of state fits to the data for various temperatures. Open circles represent Earth model PREM – the upper set is ρ, the lower set is Φ. There is relatively good agreement between the data and the extrapolated values for perovskite at the 2,000 K adiabat. The inset shows the zero-pressure dependence of thermal expansion on temperature used in the extrapolation. The shaded region separates values consistent with a perovskite lower mantle from those that are consistent with a pyrolite lower mantle. These results suggest that silicate perovskite is the dominant lower mantle phase.

is necessary to have an accurate equation of state for silicate perovskite and magnesiowüstite at high pressures, including a thermal expansion correction for the higher temperatures in the lower mantle.

Knittle et al. (1986) inferred a large thermal expansivity for the perovskite structure in the lower mantle, in the neighborhood of $4 \times 10^{-5}\,K^{-1}$, based on extrapolations of measurements made at low pressures and temperatures. With this value, Jeanloz and Knittle (1989) concluded that the pyrolite model of upper mantle composition matches the PREM density profile in the lower mantle only if the temperature is low, in the neighborhood of 1,500 K. More likely temperatures, in excess of 2,000 K, imply that the pyrolite model composition is less dense than the lower mantle, by about $4 \pm 2\%$, according to their analyses. This interpretation is far from certain, however, in part because pressure acts to reduce the thermal expansivity of perovskite, making it difficult to distinguish between it and magnesiowüstite on the basis of density alone (Chopelas and Boehler, 1992).

A more complete comparison of magnesioperovskite with lower mantle density and seismic velocity is shown in Figure 3.18, from Stixrude et al. (1992). This comparison demonstrates that a pure perovskite composition matches both the density and the bulk sound velocity of the lower mantle, at realistic lower mantle temperatures. In contrast, Stixrude et al. (1992) find that a pyrolite (olivine-rich) upper mantle composition does not fit the seismically observed profiles as well. The implication is that the lower mantle may be mostly perovskite $(Mg, Fe)SiO_3$, and therefore enriched in silica with respect to the dominantly olivine $(Mg, Fe)_2SiO_4$ upper mantle.

These types of comparisons depend critically on mineral properties such as thermal expansivity and on the temperature in the transition zone, which is also uncertain. In addition, they do not determine where in the mantle the compositional change occurs. They do, however, constrain the range of possible density increases due to compositional stratification to lie within the limits 0–4%. Unfortunately, even this narrow range encompasses a wide spectrum of dynamical behaviors. If the stratification is low, corresponding to a density increase of 2% or less, compositional stratification is not likely to prevent subducting slabs from sinking into the lower mantle. At the upper limit, where the density increase due to composition is 4%, the stabilizing effect of chemical buoyancy is as great or greater than the destabilizing effect of thermal buoyancy, and subducted slabs would not sink deep into the lower mantle. In the intermediate range, it might be possible for some slabs to sink into the lower mantle, while others might be deflected in the transition zone. The evidence from seismic tomography discussed later in this chapter indicates that some slabs do sink into the lower mantle, while others appear to be deflected. This highlights why a precise knowledge of phase and compositional changes through the transition zone is so critical in determining the structure of mantle convection.

Better resolution of the seismic discontinuities and other fine structure in the transition zone continues to be an important goal of seismology. In addition to the seismic discontinuities at 410 and 660 km depth already discussed, there is evidence for another near 520 km depth. The 520 km seismic discontinuity is seen in topside reflections following P and S (Shearer, 1990, 1991), in underside reflections preceding SS (Shearer, 1991, 1996), and in ScS reverberations (Revenaugh and Jordan, 1991). However, the 520 km seismic discontinuity has not been detected in refraction studies. All of the seismic data are consistent with a global 520 km seismic discontinuity if the feature is spread out over a depth interval of 10–50 km and has changes in seismic velocities small enough to have avoided detection in routine seismic refraction studies (Shearer, 1996). The observations of the 520 km seismic discontinuity constrain the change in seismic impedance (product of seismic velocity and density) across the discontinuity to lie between about 2 and 4%. Most of the impedance change must occur in density rather than seismic velocity to account for the lack of detection of the 520 km seismic discontinuity in refraction studies. Shearer (1996) suggests that the known properties of the β-modified spinel $\rightarrow \gamma$ spinel phase change are consistent with the seismic observations of the 520 km discontinuity.

The global topography of the 520 km discontinuity has been determined by Flanagan and Shearer (1998a); though subject to considerable uncertainty, the discontinuity appears to be elevated in the north and south central Pacific and depressed in the mid-Pacific. Flanagan and Shearer (1998a, 1999) also report on the large-scale topography of the 410 km seismic discontinuity which is lower in amplitude and largely uncorrelated with the topography in the 660 km discontinuity. In the vicinity of subduction zones there is little change from the average depth of the 410 km discontinuity, though depth variations perpendicular to the strike of the Tonga slab are as large as about 30 km and consistent with the elevation of the olivine–β-modified spinel phase boundary in cold regions (Flanagan and Shearer, 1998b).

3.6 The Lower Mantle

The lower mantle is 70% of the mass of the solid Earth and nearly half of the mass of the entire Earth. It is marked by a smooth variation in seismic properties with depth, a fact that was instrumental in the success of early spherical Earth models. Beneath the transition zone

and extending to within 300 km of the core–mantle boundary (CMB), a degree of uniformity prevails that is found nowhere else in the mantle. Early seismic array studies (Johnson, 1967) suggested small departures from isentropic conditions, in the form of subtle higher-order discontinuities. The trend of recent spherical models, however, is toward ever more homogeneity in the lower mantle, with the important exception of the layer D'' just above the core. As an illustration, Figure 3.19 shows the difference between the Bullen inhomogeneity parameter η and unity for three spherical Earth models (see also Brown and Shankland, 1981). The Bullen inhomogeneity parameter η is defined as

$$\eta \equiv \frac{K_a}{\rho} \frac{d\rho}{dp} = \Phi \frac{d\rho}{dp} \tag{3.6.1}$$

From (3.2.7) it is seen that $\eta = 1$ for isentropic conditions. Equation (3.6.1) can be rewritten using the hydrostatic relation (3.2.8) as

$$\eta = \frac{-K_a}{\rho^2 g} \frac{d\rho}{dr} = \frac{-\Phi}{\rho g} \frac{d\rho}{dr} \tag{3.6.2}$$

If we consider ρ as a function of temperature T and pressure and write

$$\frac{d\rho}{dr} = \left(\frac{\partial \rho}{\partial T}\right)_p \frac{dT}{dr} + \left(\frac{\partial \rho}{\partial p}\right)_T \frac{dp}{dr} \tag{3.6.3}$$

then (3.6.2) can be further transformed using (3.6.3), (3.2.5), (3.2.8), and thermodynamic relations from Chapter 6 ((6.6.2), (6.6.3), (6.6.33), (6.6.34), (6.7.19)) into

$$\eta - 1 = \frac{\alpha \Phi}{g} \left\{ \frac{dT}{dr} - \left(\frac{dT}{dr}\right)_a \right\} \tag{3.6.4}$$

where α is the thermal expansivity and $(dT/dr)_a$ is the adiabatic temperature gradient. For isentropic conditions it is again seen from (3.6.4) that $\eta = 1$. The quantity $\eta - 1$ is zero in a homogeneous isentropic region, it exceeds zero in regions with subadiabatic or stabilizing temperature gradients (since dT/dr is normally negative in the mantle and dT_a/dr is negative, subadiabaticity means $-dT/dr < -(dT/dr)_a$ or $dT/dr > (dT/dr)_a$ and $\eta - 1 > 0$), and it is less than zero in regions with destabilizing or superadiabatic temperature gradients (superadiabaticity means $-dT/dr > -(dT/dr)_a$ or $dT/dr < (dT/dr)_a$ and $\eta - 1 < 0$). From (3.2.6) and (3.6.1) the quantity $\eta - 1$ can also be written:

$$\eta - 1 = \frac{d\rho}{d\rho_a} - 1 \tag{3.6.5}$$

where $d\rho_a$ is the adiabatic or isentropic density change in the radius interval dr. If non-isentropic conditions are due to compositional density gradients, then stable compositional stratification requires $-d\rho > -d\rho_a$ (since $d\rho/dr$ is normally negative in the mantle and $d\rho_a/dr$ is negative) or $d\rho_a > d\rho$ and $\eta - 1 > 0$. Unstable compositional stratification requires $-d\rho < -d\rho_a$ and $\eta - 1 < 0$. Whether due to temperature or composition, $\eta - 1 > 0$ means stability against overturning and $\eta - 1 < 0$ means instability. Seismic Earth models such as PREM closely approximate the homogeneous, isentropic limit $\eta = 1$, from 800 km depth down to the D'' layer (Figure 3.19). Within the limits of seismic resolution, the spherically averaged lower mantle is essentially homogeneous and isentropic.
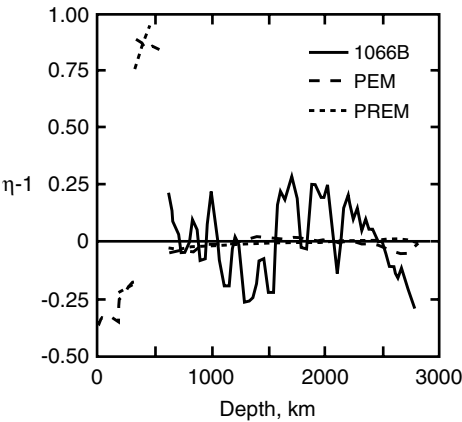
Figure 3.19. The difference between the Bullen inhomogeneity parameter η and unity through the mantle from three seismic Earth models (Shankland and Brown, 1985). The quantity $\eta - 1 = 0$ corresponds to a homogeneous, isentropic (well-mixed) state.

Why does the lower mantle appear so homogeneous in spherical Earth models? Part of the answer is the apparent absence of additional phase transformations at pressures higher than those in the transition zone. So, a mixture of silicate perovskite and magnesiowüstite can explain nearly all of the radial structure of the lower mantle. Ironically, this implies that magnesium silicate perovskite, a mineral unknown in the crust or mantle xenoliths, is the Earth's single most abundant mineral! The fact that Bullen's homogeneity index (Figure 3.19) is nearly unity through the lower mantle (apart from the D″ layer) indicates one of two possible states. The simplest state assumes uniform composition, and implies that the lower mantle has been homogenized by convective mixing. This assumption is implicit in the standard, whole mantle convection model.

The other possible state assumes that the lower mantle is nonuniform in both composition and temperature, but that their individual effects on Bullen's homogeneity index tend to cancel each other. In this state the lower mantle would appear to be grossly homogeneous, even though it is stratified. Convection that results from a close balance between thermal and chemical buoyancy forces in a fluid is called thermochemical convection. A model for large-scale thermochemical convection in the lower mantle due to chemical stratification has been proposed by Kellogg et al. (1999), as a way of reconciling the evidence of deep slab penetration discussed in this chapter with the need for long-lasting mantle isotopic reservoirs discussed in Chapter 12. Some dynamical implications of thermochemical mantle convection are discussed in Chapters 9 and 10.

3.7   The D″ Layer and the Core–Mantle Boundary

Question 3.6: *Is the D″ layer a purely thermal boundary layer or does it have a compositional component?*

A portion of the lower mantle that is definitely not homogeneous is the 300 km above the CMB, the so-called D″ layer. One of the most critical regions in the Earth in terms of geodynamical processes, it governs the interaction between the core and the mantle; it may also be the source of deep mantle plumes, a zone of active chemical reactions, and an
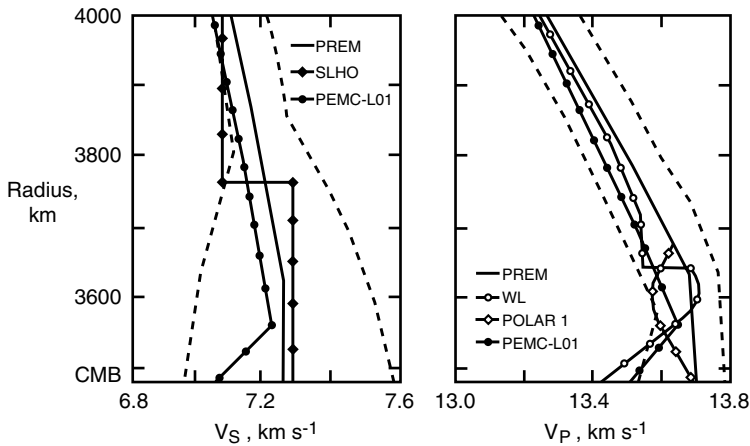
Figure 3.20. Profiles of V_S and V_P through the D″ region, including models PREM (Dziewonski and Anderson, 1981), SLHO and WL (Young and Lay, 1987), PEMC-L01 (Doornbos and Mondt, 1979), and POLAR 1 (Ruff and Helmberger, 1982). The dashed lines delineate an approximate envelope of the V_S and V_P profiles. After Lay (1989).

important geochemical reservoir (Loper and Lay, 1995). The D″ layer is now considered to contain the hot, basal thermal boundary layer of mantle convection, and seismological evidence indicates it may also be compositionally distinct from the rest of the lower mantle.

It is remarkable how much is known about the core–mantle boundary region from seismology (see Wysession et al., 1998, for a review). The early Earth models of Jeffreys, Bullen, and Gutenberg included a region 100–200 km thick above the CMB with negative P- and S-wave velocity gradients. These structures seemed to be demanded by observations of waves diffracted along the CMB. Subsequent investigations using high-frequency diffracted body waves (Doornbos and Mondt, 1979) yielded models with a slightly negative velocity gradient through the 75–100 km interval just above the CMB, as shown in Figure 3.20. More recent seismic studies (Garnero and Helmberger, 1996) have provided additional resolution of this region, and indicate the presence of a laterally variable ultra-low velocity zone (ULVZ) immediately above the CMB, with a velocity reduction of about 10% relative to the overlying mantle and a thickness of 40 km or less. The ULVZ is present beneath the central Pacific but appears to be either vanishingly thin or missing in other regions, such as beneath the Americas. Its large, negative velocity anomaly and the apparent correlation between its thickness variations and the three-dimensional structure of the lower mantle as imaged by global seismic tomography (see Figure 3.32) has led to the interpretation that this zone represents accumulation of partial melts, and is perhaps the source of deep mantle plumes (Williams and Garnero, 1996). Vidale and Hedlin (1998) have reported unusually large amplitude precursors to PKP due to scattering from a 60 km thick layer at the CMB north of Tonga. This scattering patch of the CMB lies in the "Equatorial Pacific Plume Group" structure distinguished by low S-wave velocities in the bottom few hundred kilometers of the mantle (Su et al., 1994). It is also near very low P wave velocities just at the base of the mantle (Garnero and Helmberger, 1996) which have been interpreted as due to partial melt (Williams and Garnero, 1996). Accordingly, Vidale and Hedlin (1998) explain the strong scattering at this patch as due to partial melt and suggest accompanying vigorous small-scale convection. The existence of partial melt at the base of the mantle is consistent

with data on the melting of lower mantle material; Holland and Ahrens (1997) used shock compression measurements of a perovskite and magnesiowüstite assemblage to place an upper bound on the solidus temperature of the assemblage of $4,300 \pm 270$ K at 130 ± 3 GPa (approximately the pressure at the CMB).

Question 3.7: *Is there partial melt at the base of the mantle?*

But D″ contains more than just a low-velocity zone. Evidence for additional structure in D″ first came from a set of seismic phases intermediate between the phases S and SKS, which has been identified in several regional studies by Lay and others (Lay and Helmberger, 1983; Young and Lay, 1987, 1990; Weber, 1993) and interpreted as refractions from a seismic discontinuity with a 2–3% shear velocity increase located 250–300 km above the CMB. Figure 3.21 is a compilation by Nataf and Houard (1993) indicating that this discontinuity has since been delineated over large portions of the CMB. Figure 3.20 shows that the velocity profiles inferred for D″ have either a negative gradient or a reduced positive gradient relative to the spherical Earth model PREM. In order to match absolute travel times, models with this discontinuity often contain another layer with reduced shear velocity gradient above the discontinuity.
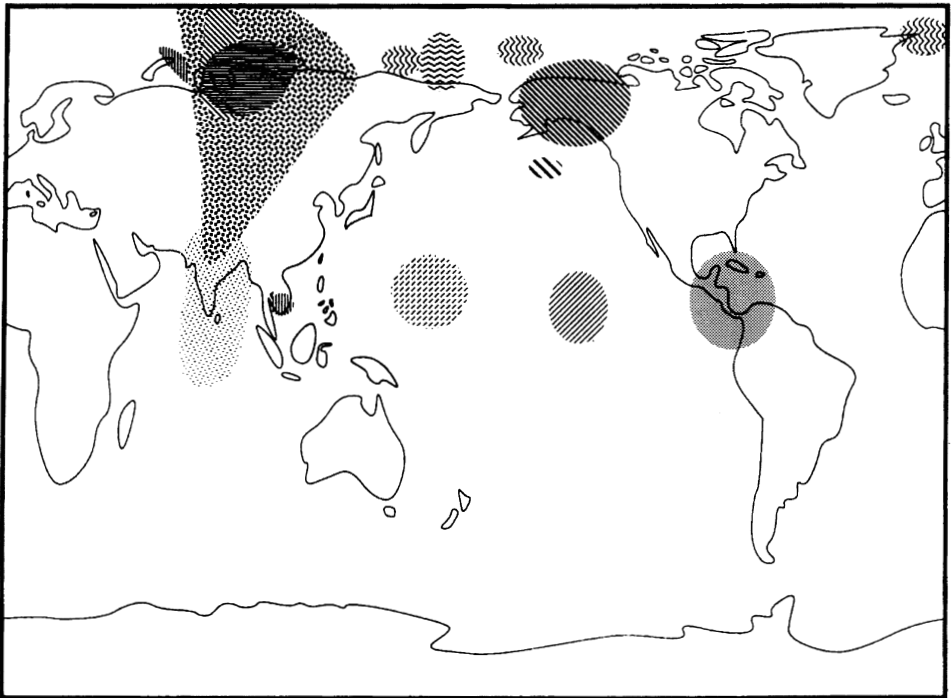

Figure 3.21. Composite map showing the different regions where a seismic discontinuity with a velocity increase has been detected at the top of *D″*. Lightly shaded regions indicate small seismic velocity anomalies in D″ and darkly shaded regions indicate large seismic velocity anomalies. From Nataf and Houard (1993).

The seismic observations suggest a composite structure for D″, consisting of a sharp velocity increase located 150–300 km above the CMB, a layer with reduced velocity gradients just beneath, and the ULVZ just above the CMB. Reduced or even negative radial seismic velocity gradients are likely to be the result of the strongly superadiabatic temperature distribution in D″. This interpretation is supported by evidence for high seismic attenuation in D″ relative to the rest of the lower mantle (Shore, 1984). Low seismic velocities and high attenuation indicate that D″ is analogous to the LID–LVZ structure – a region with steep thermal gradients and perhaps partial melting. D″ and the LID–LVZ are, respectively, hot and cold thermal boundary layers produced by large-scale mantle convection (Verhoogen, 1973; Jones, 1977; Elsasser et al., 1979). However, some of the structure in D″ is likely to be compositional in origin (Loper and Lay, 1995). The discontinuity at the top of D″ might indicate a change in bulk composition, or alternatively it might represent an as yet unexplained pressure-induced phase change.

Global-scale variations in the thickness of the D″ layer are evident from the comparison of regional studies shown in Figure 3.20. In addition, there is also seismic evidence for heterogeneities in D″ distributed over shorter length scales. Scattering of high-frequency P waves indicates the presence of either roughness on the CMB or small-scale heterogeneities distributed within D″ (Haddon, 1982; Haddon and Buchbinder, 1987; Bataille and Flatté, 1988). The estimated wavelengths of the scatterers are in the range 20–200 km.

In addition to fine-scale heterogeneity in D″, long-wavelength topographic undulations of the CMB and variations of the thickness of D″ are expected. Broad-scale undulations of D″ and the CMB probably represent dynamic topography supported by stresses in the mantle that drive the convective motion. Dynamic calculations predict that mantle convection supports variations of several kilometers on the CMB (Hager and Richards, 1989; Forte et al., 1993). There have been numerous attempts to map dynamic topography on the CMB using the techniques of seismic tomography; these will be discussed later in this chapter.

The significance of the broad spectrum of heterogeneity in D″ and CMB topography derives from the fact that these cannot be static features. They must be dynamically supported by nonhydrostatic stresses in the lower mantle. Unsupported CMB topography would be expected to relax on a short time scale considering the rheology at the high temperatures in D″. Since the lifetime of unsupported topography is short, such topography must be supported dynamically and this implies a close connection between the pattern and amplitude of CMB and D″ undulations and the pattern of flow in the lower mantle, both at short and long wavelengths. The fact that heterogeneity is present on many scales indicates that the dynamical processes in D″ are multi-scale in character, and involve both thermal and compositional effects.

3.8 The Core

Seismic evidence for the outer core was first provided by Oldham (1906), who interpreted the shadow zone between epicentral angles 100° and 120° as a low-velocity shell from the Earth's center out to about 3,000 km radius. Gutenberg (1913) estimated the core radius at 3,471 km, within 0.5% of the presently accepted value 3,485 ± 3 km (Engdahl and Johnson, 1974). The inner core was discovered by Lehmann (1936), who recognized that seismic compressional waves traversing the inner core, the phase PKIKP, require a discontinuous increase in V_P near 1,220 km radius. Bullen (1946) interpreted this increase as the effect of finite rigidity, implying that the inner core is solid. The mean radius of the inner core is currently estimated at 1,216 ± 3 km (Souriau and Souriau, 1989).

Liquidity of the outer core was first established using solid Earth tide amplitudes and the period of the Chandler wobble (Jeffreys, 1929) and has been substantiated by the complete lack of evidence for shear wave transmission. However, the direct demonstration of inner core rigidity – an observation of shear wave propagation through the inner core via the phase PKJKP – is still controversial. The inner core rigidity indicated by the nonzero V_S in Figure 3.2 is inferred from normal mode frequencies that are particularly sensitive to core structure. It is interesting that the modern estimate for the inner core seismic shear velocity is within 10% of Bullen's (1946) original estimate. Poisson's ratio is evidently about 0.4 in the inner core, which is quite large in comparison with 0.24–0.26 found elsewhere in the solid Earth. For an ideal solid, with $V_P = V_S \sqrt{3}$, Poisson's ratio is 0.25, and is 0.5 in the liquid outer core. Metals typically have large Poisson ratios at high pressure, another indication of the metallic nature of the core.

Within the resolution of spherical Earth models, both the outer core and the inner core appear radially homogeneous, except perhaps in the neighborhood of the inner core boundary (ICB), where the inhomogeneity parameter η deviates from unity. Both of the core boundaries appear to be sharp. The CMB is no wider than 10 km (Engdahl et al., 1974) and the ICB is even sharper – probably 3 km or less. The density jump at the ICB, of great significance for core composition and for powering the geodynamo, has been estimated by Masters and Shearer (1990) to be $550 \pm 150\,\mathrm{kg\,m^{-3}}$.

At seismic frequencies, the outer core shows practically no damping. Qamar and Eisenberg (1974) have estimated $Q \simeq 5,000$ from P waves multiply reflected off the underside of the CMB, while Cormier and Richards (1976) argue for an even larger value. The inner core is quite different in that respect. Doornbos (1974) obtained a Q profile from 1 Hz P waves that increases from 200 just beneath the ICB to near 2,000 at the Earth's center. Attenuation of radial normal modes requires a Q of about 200 near the ICB, rising to near 1,000 at the Earth's center (Masters and Shearer, 1990). The inner core is also seismically anisotropic, and this property has led to a discovery with important implications for the operation of the geodynamo. Seismic waves propagating through the inner core travel a few percent faster along the north–south direction than along paths through the equator (Poupinet et al., 1983; Souriau and Romanowicz, 1997). When this difference was first recognized, the initial interpretation was that the inner core is transversely anisotropic, with a seismically fast axis aligned with the Earth's rotation axis (Morelli et al., 1986; Shearer et al., 1988; Creager, 1992; Song and Helmberger, 1993). Subsequently, Su and Dziewonski (1995) showed that a better fit to the data is obtained by allowing the axis of anisotropy to be inclined to the rotation axis by about 10°, although Souriau et al. (1997) question the reliability of the tilt determination. Explanations for the cause of inner core anisotropy include crystal alignment due to freezing controlled by the magnetic field (Karato, 1993), a simple large-scale inner core convection pattern (Romanowicz et al., 1996), or stress-induced texture accompanying heterogeneous growth (Yoshida et al., 1996).

The small deviation of the seismically fast axis from rotational alignment provides an explanation for reports of secular variation in travel times of seismic waves traversing the inner core (Song and Richards, 1996). Rotation of the inner core relative to the crust and mantle would change the inner core velocity structure along a given ray path, so that the travel times of waves from earthquakes originating from the same focal region and recorded at the same station would change with time. On this basis, Song and Richards (1996) inferred a $1°\,\mathrm{yr^{-1}}$ prograde (eastward) anomalous rotation of the inner core and Su et al. (1996) inferred an even larger inner core prograde rotation, nearly $3°\,\mathrm{yr^{-1}}$. However, re-examination of the problem by Souriau et al. (1997) and by Laske

and Masters (1999) suggests caution in accepting the reality of inner core differential rotation.

The possibility that the inner core might rotate at a slightly different rate than the crust and mantle was first pointed out by Gubbins (1981), who argued that the low viscosity of the outer core fluid decouples the inner core from the silicate portion of the Earth, and that electromagnetic torques applied to the inner core can easily support anomalous rotation. One of the striking features of the convective dynamo simulations of Glatzmaier and Roberts (1995, 1996) is a prediction of anomalous prograde inner core rotation, at approximately one degree per year. In the Glatzmaier–Roberts dynamo simulations, strong prograde (eastward) flow occurs in the outer core fluid adjacent to the ICB, driven by large-scale density variations. The inner core responds to flow in the outer core through electromagnetic coupling (Aurnou et al., 1996) and also responds to mantle heterogeneity through gravitational coupling (Buffett and Glatzmaier, 2000). Thus, the inner core motion is a sensitive indicator of the torque balance near the Earth's center.

Recent analyses of PKP phases traveling along east–west paths have revealed more detail about seismic heterogeneity and anisotropy in the inner core (Tanaka and Hamaguchi, 1997). P-wave velocities in the outermost 100–500 km of the inner core show a degree 1 heterogeneity with relatively fast velocities in an "eastern" hemisphere and slow velocities in a "western" hemisphere. Analysis of PKP phases traveling nearly parallel to the Earth's spin axis revealed significant anisotropy only in the "western" hemisphere. The east–west hemispheric asymmetry in P-wave velocity and seismic anisotropy may be related to core dynamics, i.e., patterns of convection in the outer core, during inner core solidification (Tanaka and Hamaguchi, 1997). Hemispheric differences in the seismic structure of the inner core were also found by Shearer and Toy (1991) and Creager (1992).

The average density in the core, approximately $11,000 \, \text{kg m}^{-3}$, greatly exceeds the average mantle density, $4,500 \, \text{kg m}^{-3}$, making it implausible that the core could be a high-pressure phase of the mantle. Yet it was not until 1963 that Birch (1963), using velocity–density systematics to show that the mean atomic weight of the core is only slightly less than the atomic weight of iron, demonstrated the core and mantle to be compositionally distinct.

The outer core is certainly an alloy of iron and lighter compounds. The inner core is richer in iron than the outer core, although it probably contains lighter alloys as well. Nickel has nearly the same physical properties as iron at high pressures (Mao et al., 1990), and consequently we have no way of determining the nickel content of the core. For purposes of discussion, we use the term "iron" to mean an iron-rich, nickel–iron alloy.

Lighter elements that could be abundant in the core include oxygen, silicon, and sulfur (Knittle and Jeanloz, 1986; Poirier, 1994a). Amounts of these elements that would individually satisfy the core density are 7, 21, and 11%, respectively. Silicon is the least likely among these, because in elemental form it is so far out of equilibrium with magnesiowüstite that it should have been absorbed into the lower mantle during core formation. The possibility of MgO solubility in liquid iron at core conditions was suggested by Alder (1966) but has not been pursued further. Thus, by the process of elimination, only oxygen and sulfur are now considered probable candidates. The likely compounds, in addition to iron–nickel, are FeO, FeS, and perhaps FeS_2.

Question 3.8: *What is the composition of the core? What is the major light element in the core?*

The energetics of the core and the consequences for mantle convection are described by Verhoogen (1980). As discussed in Chapter 13, all portions of the Earth, including the core, cool as the radioactive elements within the mantle decay. The ICB is associated with the liquid–solid phase transition for iron with the appropriate alloying elements. As the core cools, the pressure at which the liquid–solid phase change takes place decreases. Thus the ICB moves outwards and liquid iron in the outer core freezes and becomes part of the inner core. The liquid-to-solid phase change is exothermic so that heat is generated upon freezing. For each degree Kelvin of cooling the heat liberated by the phase change is approximately equal to the loss of heat associated with the cooling (Loper, 1978a; Stevenson, 1981). The rate of crystallization is about $10^6 \, \mathrm{kg \, s^{-1}}$ if one assumes the inner core has grown steadily to its present size in 4 Gyr.

Braginsky (1963) was the first to emphasize how the release of potential energy accompanying inner core freezing could provide a power source for the geodynamo. As iron-rich material freezes onto the inner core, the rejected liquid is enriched in light components, and is buoyant with respect to the outer core liquid. Accordingly, it must rise and mix with the outer core. This motion, called thermochemical convection, is likely to be a dominant source of kinetic energy powering the dynamo. It has been shown theoretically that thermochemical convection is more efficient at dynamo action than purely thermal convection, in the sense that it requires a smaller heat loss from the core to produce the same amount of magnetic field energy (Loper, 1978a; Verhoogen, 1980). It is estimated that only about 5 TW of core heat loss is needed to sustain the dynamo by thermochemical convection and that compositional buoyancy provides 2–4 times more power to the dynamo than does thermal buoyancy (Cardin and Olson, 1992; Lister and Buffett, 1995).

The relative importance of thermal to compositional convection in the core depends on the magnitude of the density increase at the ICB that is due to the change in composition. The part of the density increase due to the phase change from liquid to solid was estimated by Verhoogen (1961) to be less than $300 \, \mathrm{kg \, m^{-3}}$, but this value is highly uncertain. The remainder, $250 \pm 150 \, \mathrm{kg \, m^{-3}}$, is evidently due to iron enrichment. This latter part is available to drive compositional convection.

Question 3.9: *What is the relative importance of thermal and compositional convection in powering the geodynamo?*

Direct information on the composition of the inner core and the outer core has come from high-pressure equation of state measurements on iron compounds at core pressures. Figure 3.22 shows the 300 K density of solid iron in the ε (hexagonal) phase versus pressure, as determined by Mao et al. (1990) from both static (diamond cell) and dynamic (shock wave) measurements. Also shown are density versus pressure curves for the Earth model PREM. Inner core density is less than the density of solid ε-iron at 300 K, and it is still less than the PREM density when the thermal expansion correction is applied. The most plausible explanation for the discrepancy is that the inner core also contains some of the light alloy present in the outer core. Jephcoat and Olson (1987) concluded that about 5% sulfur in the inner core and 10–12% in the outer core matches the PREM density profile. Alternatively, an inner core with about 3% oxygen and an outer core with 7–8% oxygen matches the PREM density equally well (Poirier, 1994a).
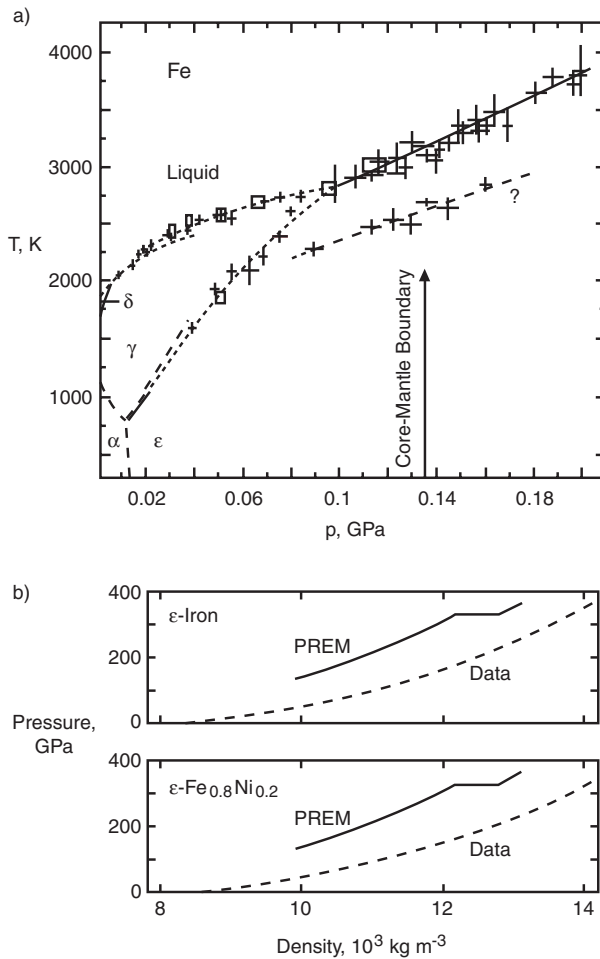
Figure 3.22. Equation of state data on iron at high temperature and pressure. (a) Experimentally determined phase diagram of iron, from Boehler (1993). The ε phase (hcp) appears to be the stable solid phase at core pressures. (b) Extrapolation of experimental pressure–density data for ε-iron and an iron–nickel alloy compared with Earth model PREM through the core, from Mao et al. (1990). Both the outer core and the inner core are systematically less dense than either metal. This discrepancy exists even after temperature corrections are applied, indicating that the core contains some lighter compounds.

3.9 Three-dimensional Structure of the Mantle

Until rather recently, global seismology was concerned almost entirely with the Earth's radial structure. Lateral variations in seismic properties were recognized, but except for the upper 200 km or so, there was no systematic effort to resolve them on a global scale. Above 200 km depth there is a good correlation between seismic structure and large-scale tectonics, which is generally associated with variations in the structure of the lithosphere. For example, the LID is thickest beneath shields and platforms on continents and is practically nonexistent beneath spreading centers and continental regions with active extensional tectonics. The oceanic LID thickness varies with the square root of crustal age, as predicted by thermal boundary layer models of lithospheric cooling. All of this structure was anticipated on the basis of models

derived from other geophysical observations. Seismologists have now gone beyond this stage, and have provided three-dimensional images of the whole mantle. These images are obtained by a variety of techniques, collectively referred to as seismic tomography.

Lateral variations in seismic wave speeds must be related to the pattern of flow in the mantle. Horizontal (isobaric) differences in V_P and V_S result from differences in the elastic constants and density. These differences can be related to the horizontal gradients in density that drive mantle flows. In principle, well-resolved three-dimensional seismic images can provide important constraints on the body forces driving mantle convection. In practice, most of the interpretation of seismic tomographic images is qualitative and consists of a search for correlations with geological structure and other geophysical observables. Quantitative uses of seismic tomographic results have been made in some instances (see the discussion of mantle rheology in Chapter 5), but a number of simplifying assumptions must be made, and the validity of these assumptions has not been fully established.

The first limitation of seismic tomography is spatial resolution. Consider, for example, the problem of resolving the structure of subducted slabs located deep in the mantle, certainly one of the primary goals of this technique. The current resolution in global tomographic models is typically 500 km horizontally and 100 km vertically. Thus, velocity anomalies associated with deep slabs are likely to be smeared out.

Limits in lateral resolution affect the identification of structures directly related to the convection process. We know from the structure of subducted slabs that much of the lateral temperature heterogeneity in the upper mantle, and perhaps everywhere in the mantle, is concentrated in zones typically 100 km wide, the characteristic thickness of the oceanic lithosphere from which slabs are derived. Clearly, narrow slab structures are at the lower limit of what can be accurately resolved by global tomography. The situation for subducted slabs is improved somewhat by the fact that the earthquake density is high in the seismogenic part of slabs so that the resolution of fine structure around the Wadati–Benioff zones is better than for the mantle as a whole (van der Hilst et al., 1991). Still, some of the structures that are important in mantle dynamics may be missing from global tomographic models.

Another difficulty in interpreting seismic tomographic images is due to the fact that isobaric variations in seismic velocity are not simply related to density variations. Instead, they are sensitive to differences in composition and temperature, and to changes in phase (both solid–solid and solid–liquid), which are themselves dependent on composition and temperature. The problem is compounded when seismic velocity anomalies at different depths are compared, because of the sensitivity of V_P and V_S to changes in pressure. In short, quantitative applications of tomographic images involve uncertainties. However, this does not diminish the importance of the qualitative interpretation of tomographic images. Indeed, they are now one of the most compelling lines of evidence for the internal structure of mantle convection.

The data used in global seismic tomography include body wave travel times and wave-forms, surface wave dispersion, and normal mode frequencies. The current generation of three-dimensional models has at its disposal several million travel times plus surface wave dispersion and waveform data over several thousand paths. Two different discretizations of the mantle are used. One divides the mantle into three-dimensional elements (boxes), and within each element, the perturbation of V_P or V_S from a spherical Earth model is calculated, based on the anomalous travel time or amplitudes along rays passing through the element. This technique was first used by Clayton and Comer (1983) and is the basis of the high-resolution models by Grand (1987, 1994) and by van der Hilst et al. (1991, 1997). The approach has also been used by Vasco and Johnson (1998), who employ arrival

times of 10 compressional and shear phases and two sets of differential times to infer the three-dimensional seismic velocity structure of the whole Earth from crust to inner core.

Alternatively, the anomalous seismic velocity distribution can be expressed as a truncated sum of spherical harmonics and orthogonal radial functions, as developed by Dziewonski (1984) and Woodhouse and Dziewonski (1984). The inversion procedure consists of solving for the coefficients of the radial functions for each spherical harmonic. This approach has been used to construct numerous global models of the entire mantle (Tanimoto, 1990; Su et al., 1994; Li and Romanowicz, 1996; Masters et al., 1996, among others).

Each approach has its merits and drawbacks. The element method computes the model only where there are sufficient data; however, the model is generally not complete and smooth. Truncated spherical harmonic representations result in a smooth and continuous model, but the model reliability is variable, particularly where the data are sparse. Normal modes have been used in two ways. Normal modes are split into $2l + 1$ separate frequencies (singlets) at each spherical harmonic degree l. The splitting is partly due to the Earth's rotation, through the Coriolis acceleration and the ellipticity. These effects can be calculated and their contribution removed from the data, leaving the splitting due to lateral structure. Only for the lowest degree modes has the splitting been measured, but they are a particularly useful data set because they are unbiased by baseline (timing) errors or uneven coverage (Dziewonski and Woodhouse, 1987). At larger spherical harmonic degrees splitting cannot be resolved, but there is a frequency shift relative to the frequency expected from a radially symmetric Earth model. Masters et al. (1982) used measurements of the frequency shift from fundamental mode Rayleigh wave spectra to deduce the presence of degree $l = 2$ heterogeneity.

Finally, long period surface and body wave data can be used in the time domain, by comparing seismograms from earthquakes with known source mechanisms to synthetic seismograms computed using radially symmetric Earth models. This is called the waveform method. None of these methods resolve the structure uniformly throughout the mantle. P-wave travel time anomalies are less useful for global studies of the upper mantle, and conversely, surface waves do not sample the lower mantle. Normal mode inversions discard the phase information and recover only the structure at even spherical harmonic degrees. Shear waves have proven to be particularly well suited to image the lower mantle, since their travel time anomalies are large.

Reviews of results of mantle tomography have been given by Woodhouse and Dziewonski (1989), Masters (1989), Hager and Clayton (1989), Romanowicz (1991), Montagner (1994), Ritzwoller and Lavely (1995), and Kennett and van der Hilst (1998). Su and Dziewonski (1991) argue that long-wavelength heterogeneities are dominant in the mantle.

3.9.1 Upper Mantle Seismic Heterogeneity and Anisotropy

Figure 3.23 shows the perturbation in shear wave velocity at a depth of 210 km, relative to the spherical model PREM, obtained by Su et al. (1994). The range of velocity perturbations is quite large, $\pm 3\%$. The correlation of low shear wave velocity with the worldwide ridge system and the correlation of high shear wave velocity with continental shield and platform areas are both excellent. High shear wave velocities also occur beneath the oldest portions of the ocean basins. The depth extent of high-velocity roots beneath stable continental interiors and the extent of low-velocity roots beneath spreading centers can be inferred from the sequence of images at different upper mantle depths shown in Figure 3.24, from the S-wave model of Li and Romanowicz (1996). High shear wave velocity beneath the
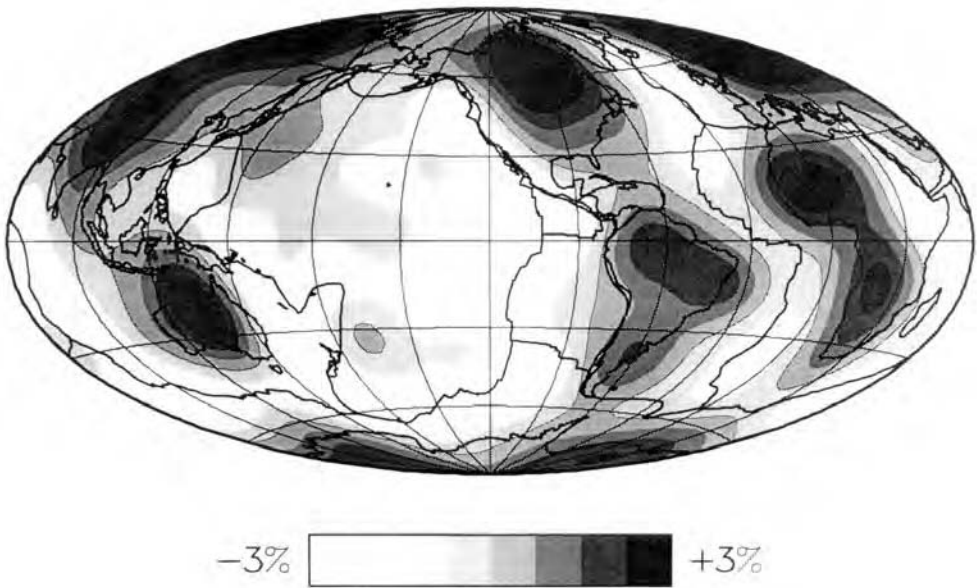
Figure 3.23. Perturbations in shear wave velocity V_S relative to the spherical Earth model PREM at a depth of 210 km complete to spherical harmonic degree 12, from the tomography model of Su et al. (1994).

Canadian shield and other continental shields is recognizable to 350 km depth. Low shear wave velocity beneath the global mid-ocean ridge system is nearly universal to 250 km depth and beneath the East Pacific, mid-Atlantic, and Antarctic ridges it appears to extend to 350 km depth.

The most important general finding from upper mantle seismic tomography is that the structure beneath the lithosphere is entirely compatible with a pattern of flow in the upper mantle constrained by plate motions. This establishes a direct connection between plate tectonics and the pattern of mantle convection. Ridges are associated with seismically slow mantle to a depth of at least several hundred kilometers. Seismic velocities beneath ridges are probably too low to be explained by excess temperature alone. More likely, there is a contribution from partial melting. The depth extent of the seismically slow region is not consistent with the upwelling of a completely adiabatic mantle. Some excess mantle temperature in the ridge root is required. However, the seismically slow region does not extend to great depths.

> **Question 3.10:** *How deep do the upwellings beneath mid-ocean ridges extend into the mantle? What are the implications for the passivity of mid-ocean ridges?*

The high seismic velocities beneath stable continental interiors evident in Figures 3.23 and 3.24 offer a partial confirmation of Jordan's (1975, 1981) tectosphere concept. If these high-velocity roots also have high strength compared to other parts of the mantle at the same depth, they imply a relatively thick continental lithosphere. The global system of subduction zones, where positive shear wave velocity anomalies of 4% are known to exist, is not well resolved in Figure 3.24. Structures with lateral dimensions of the order of 100 km or less are
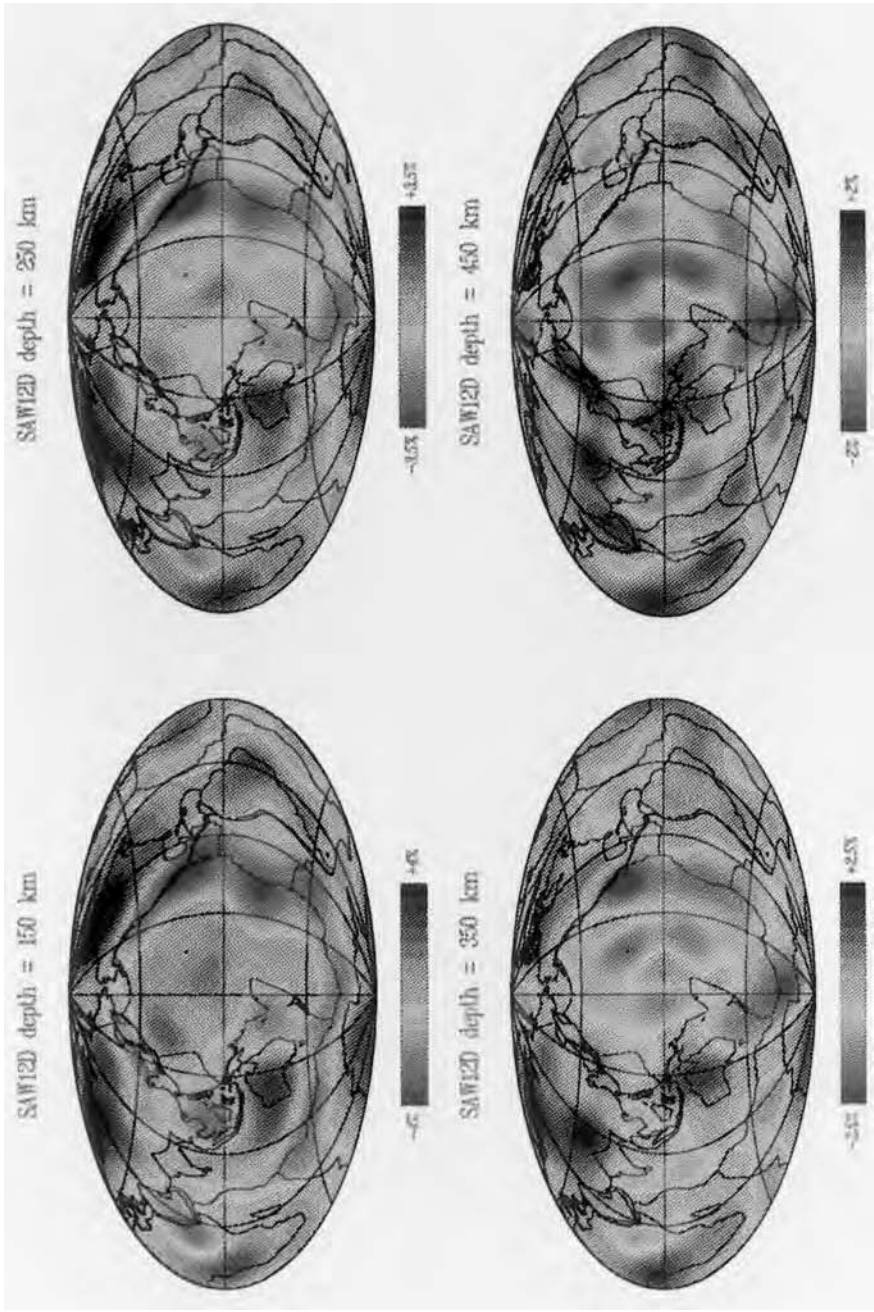
Figure 3.24. Perturbations in shear wave velocity V_S relative to the spherical Earth model PREM at various upper mantle depths from the model of Li and Romanowicz (1996).

For a color version of this figure, see plate section.
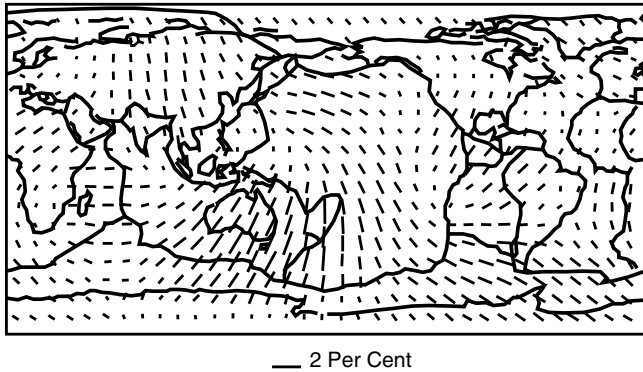
105

— 2 Per Cent

Figure 3.25. Directions of fastest principal axes beneath the lithosphere in azimuthally anisotropic Rayleigh wave propagation at 200 s period (Tanimoto and Anderson, 1985).

not resolved in global models, even if the associated seismic velocity anomalies are large. In subduction zone environments high-velocity slabs are in close proximity to low velocities in the back-arc region. This compounds the difficulty of resolving slabs with the spherical harmonic approach.

Seismic anisotropy exists in the upper mantle at about the 2% level, as discussed earlier in this chapter, and it provides information on the large-scale pattern of deformation. Lithospheric seismic anisotropy records the fossil spreading direction, whereas asthenospheric seismic anisotropy is thought to indicate the present direction of horizontal velocity, with vertical shear providing the mechanism for olivine grain alignment. Global inversions of long period surface wave data (Tanimoto and Anderson, 1985; Nataf et al., 1986) reveal a large-scale pattern of seismic anisotropy beneath the lithosphere that is generally consistent with the motion constrained by the plates (Figure 3.25). The largest seismic anisotropy occurs beneath the fastest moving plates, and there is an indication in the overall pattern of seismic anisotropy of divergence beneath the East Pacific Rise and convergence beneath the trenches in the western Pacific.

3.9.2   Extensions of Subducted Slabs into the Lower Mantle

Question 3.11: *Do subducted slabs penetrate into the lower mantle?*

The ultimate fate of the subducted lithosphere, and its trajectory beyond the seismicity cutoff, is one of the most important unsolved problems in mantle convection. The traditional interpretation of the pattern of deep seismicity is that slabs remain in the upper mantle. Early studies of focal depths indicated there are no earthquakes deeper than 700 km depth (Isacks et al., 1968; Isacks and Molnar, 1971). More recent focal depth determinations have confirmed this, and indeed the maximum depth is now judged to be 670–680 km (Stark and Frohlich, 1985). Focal mechanisms of very deep earthquakes, those just above the seismicity cutoff, mostly show down-dip compression, an indication that slabs encounter increased resistance at that depth. Evidence in support of this interpretation has been presented by Giardini and Woodhouse (1984), who detect a kink in the Tonga slab just above the 660 km seismic discontinuity. The increased resistance to subduction in the transition zone could be

due to compositional buoyancy or phase-change buoyancy associated with the endothermic phase transition. A significant increase in mantle viscosity associated with the 660 km phase change (see Chapter 5) would also result in resistance to subduction. The important difference between these explanations is that increased viscosity would deform slabs, but would not prevent them from sinking into the lower mantle, whereas compositional or phase-change buoyancy might.

Low slab temperatures result in seismic wave velocities 5–10% higher than in the surrounding mantle. Seismologists have long used this property to detect the presence of slab-related heterogeneities. An early study by Engdahl and Gubbins (1987) demonstrated that the Aleutian slab anomaly extends below the deepest seismicity, and Spakman (1986) showed that the same is true for the Hellenic slab.

Extensive evidence for aseismic extensions of slabs deeper than the 660 km seismic discontinuity in the western Pacific was assembled by Jordan (1977) and Creager and Jordan (1984, 1986a) using travel time residuals from deep focus slab earthquakes. Slab extensions below the deepest earthquakes produce a characteristic pattern of travel time anomalies. Creager and Jordan (1984, 1986a) concluded that the pattern of these travel time anomalies from deep earthquakes in the western Pacific could be explained by slab-shaped regions of high-velocity material extending to 1,400 km depth or more, with the same strike as the Wadati–Benioff zone from which the earthquakes originated. They argued that these residual sphere patterns indicate thermal anomalies from subducted lithospheric slabs extending below the transition zone and into the lower mantle. This interpretation was questioned, however, on the grounds that the effects of seismic anisotropy were not included (Anderson, 1987a) and that the travel time anomalies from deep earthquakes are contaminated by seismic heterogeneity unrelated to slabs (Zhou and Anderson, 1989).

There have been a growing number of high-resolution seismic tomographic studies attempting to answer the critical question of how slabs are affected by the 660 km seismic discontinuity and whether lithospheric material penetrates into the lower mantle (van der Hilst et al., 1991; Fukao et al., 1992; Spakman et al., 1993; Grand, 1994; van der Hilst, 1995; Grand et al., 1997; van der Hilst et al., 1997). In general, two different modes of slab interaction with the 660 km seismic discontinuity are inferred. In some cases the slab appears to descend into the lower mantle without significant change in dip angle. Examples of this behavior include the northern Kurile, Mariana, Kermadec, and Indonesia slabs. In other cases the slab appears to be bent at the 660 km seismic discontinuity. Some of the bent slabs appear to be deflected nearly into the horizontal, so they seem to be reclined on the 660 km seismic discontinuity. Examples of this behavior include the Izu-Bonin, southern Kurile, and Japan slabs. For other bent slabs the 660 km seismic discontinuity simply appears to kink the slab but not prevent it from sinking into the lower mantle. The central portion of the Tonga slab is an example of this behavior.

Regional-scale seismic tomography is particularly well suited for imaging slabs. Figure 3.26 shows a cross-section through the upper mantle beneath the Hellenic arc with the S-wave velocity structure determined by Spakman et al. (1993). The image contains a high-velocity, dipping tabular anomaly, extending to a 660 km depth beneath the arc, with low velocities in the overlying wedge. The structure is consistent with low temperatures in the subducting slab adjacent to a high-temperature back-arc region. There is a subhorizontal portion of the high-velocity anomaly located just above the 660 km discontinuity, suggesting that a kink exists in the slab at that depth. But there is also high velocity beneath the kink, an indication that the slab material continues beneath the transition zone into the lower mantle. Further evidence of this behavior is shown in Figure 3.27, from a tomographic study of
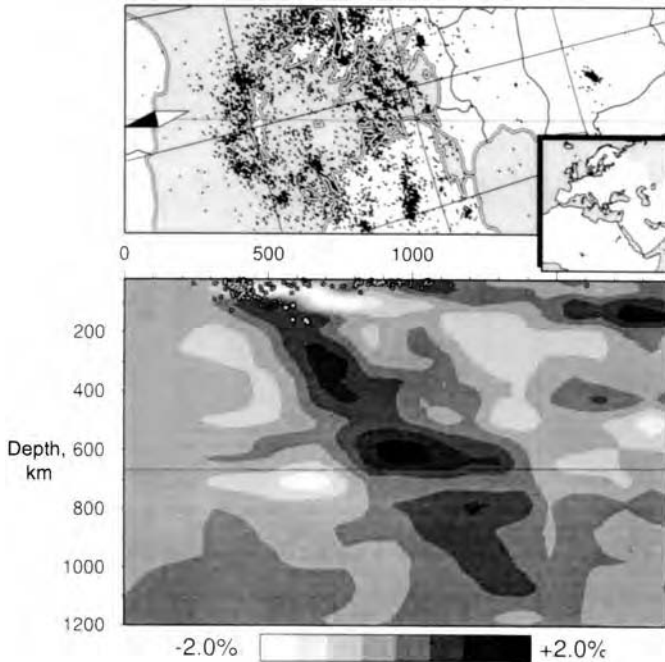
Figure 3.26. Perturbations in shear wave velocity V_S relative to the Jeffreys–Bullen spherical Earth model on a great circle cross-section through the Aegean upper mantle to a depth of 1,200 km as determined by Spakman et al. (1993). The upper panel is a reference map showing seismicity and the location of the cross-section. The horizontal dimension in the lower panel is distance in kilometers.

the mantle around the Tonga–Kermadec subduction zone by van der Hilst (1995). These images indicate along-strike variation in deep subduction, in which the slab in the northern cross-sections is deflected at 660 km depth while in the southern cross-section it appears to penetrate into the lower mantle with little change in dip angle.

Only recently has it become possible to construct global maps of mantle heterogeneity beneath the transition zone with sufficient spatial resolution to test the slab penetration hypothesis worldwide. Figure 3.28 shows a high-resolution block model P-wave tomographic image at 1,300 km depth by Widiyantoro and van der Hilst (1996) with about 90% spatial coverage. Prominent, linear high-velocity belts are present beneath the southern Asia convergence zone, west of the present-day location of the Tonga trench, and the Americas. All three of these belts coincide with the location of major convergent plate boundary regions active in the last 100 Myr (see Chapter 2). This correlation offers compelling evidence that the belts of high-velocity anomalies at 1,300 km depth represent remanents of subducted oceanic lithosphere. Additional evidence in support of this interpretation is found from the high-resolution studies by Grand (1987, 1994) of the seismic anomaly structure beneath the Americas, particularly the so-called Caribbean anomaly. Figure 3.29 shows the shear wave velocity structure beneath the Americas at 1,300–1,450 km depth determined by Grand (1994), compared with the location of the subduction zone off the west coast of North America between 60 and 90 Myr ago, as inferred from plate reconstructions by Engebretson et al. (1992). The shape and location of the Caribbean anomaly correlates well with the expected shape and location of slab material sinking in the lower mantle at 10–20 mm yr^{-1}.
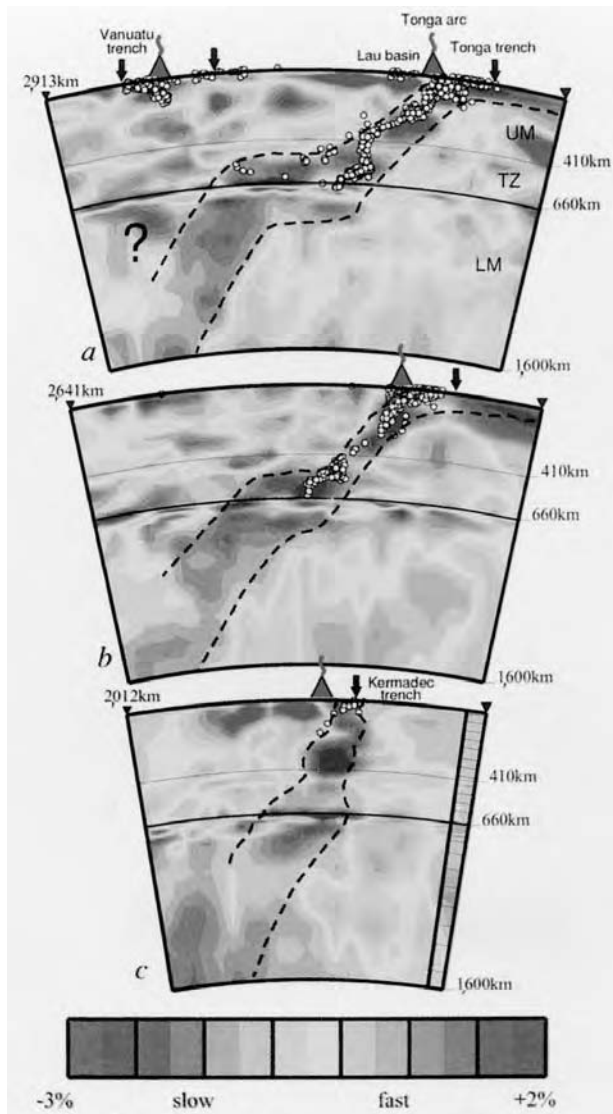
Figure 3.27. Tomographic images of seismic wave velocity V_P variations in the mantle beneath (a) the northern portion of the Tonga trench, (b) central Tonga, and (c) the Kermadec trench. Dots indicate earthquake foci. From van der Hilst (1995).

For a color version of this figure, see plate section.

The linear high seismic velocity anomalies in Figures 3.28 and 3.29 extend even deeper into the lower mantle as seen in Figure 3.30, which compares high-resolution P- and S-wave seismic tomographic models at a number of depths through the lower mantle (Grand et al., 1997). The linear anomalies beneath North and Central America and southern Asia are apparent even at 1,800 km depth. The generally good agreement between the P- and S-wave models argues convincingly for the reality of the linear high seismic velocity anomalies and their interpretation as subducted slabs.
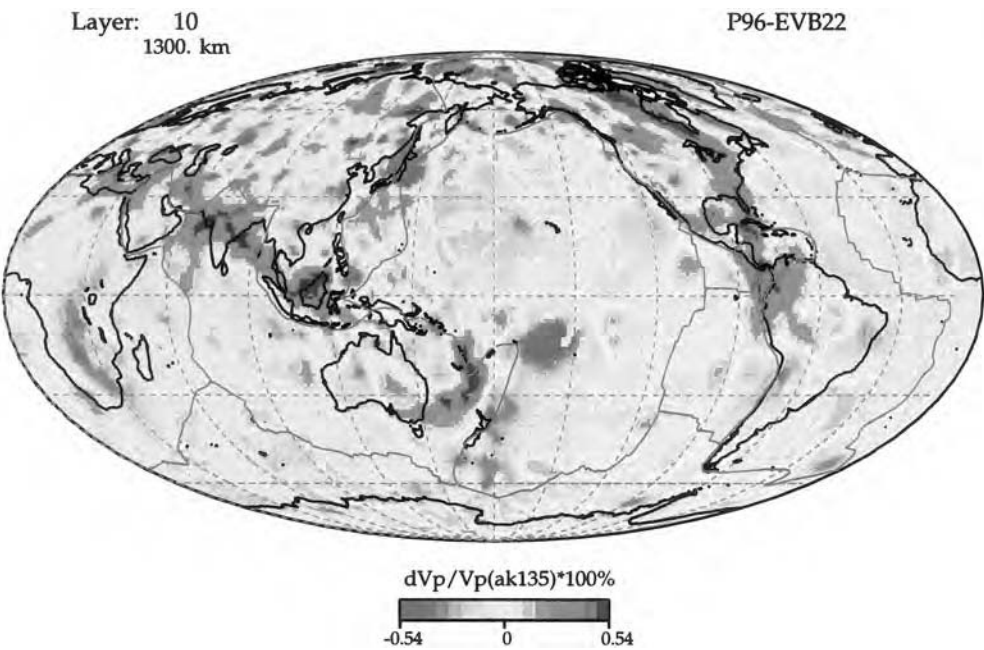
Figure 3.28. Perturbations in compressional wave velocity V_P relative to the spherical Earth model PREM at a depth of 1,300 km from the tomography model of Widiyantoro and van der Hilst (1996).

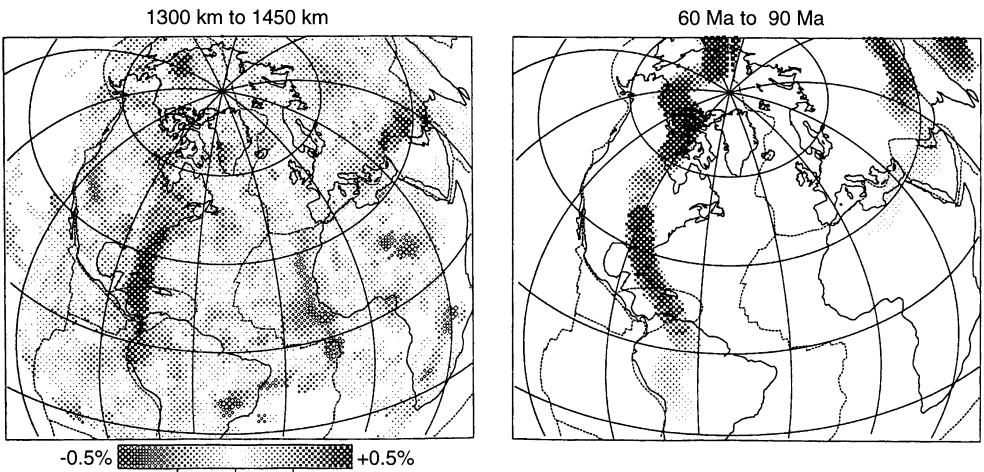*For a color version of this figure, see plate section.*

Figure 3.29. Perturbations in shear wave velocity V_S between 1,300 and 1,450 km depth beneath North America, determined by Grand (1994) (left), compared with the location of subduction zones between 60 and 90 Ma inferred by Engebretson et al. (1992) (right).
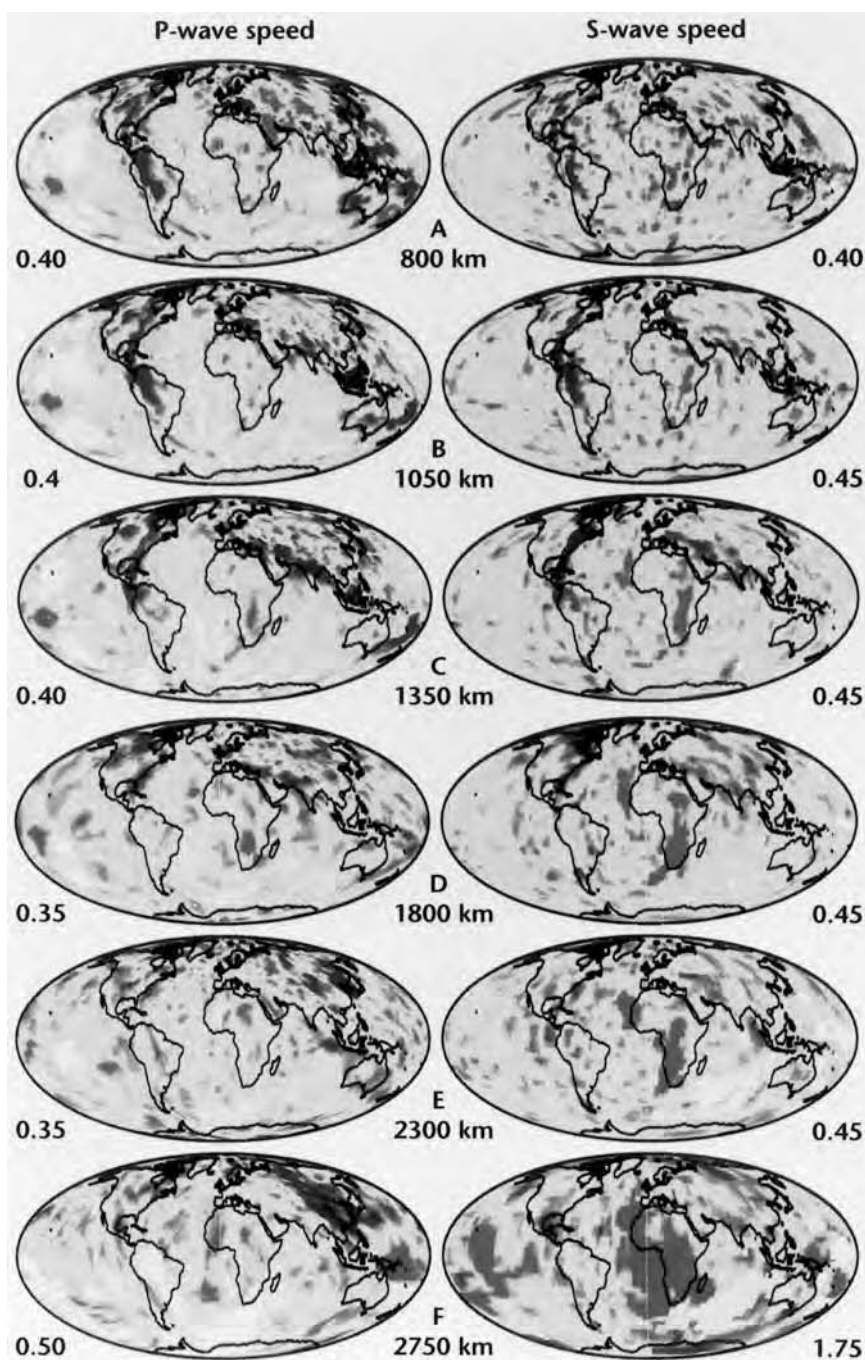
Figure 3.30. High-resolution P-wave and S-wave tomographic models of the lower mantle. Numbers at the sides of the images are the maximum anomaly in terms of percentage difference from the mean velocity. Blues are faster than average speed, reds are slower. White regions have no significant data sampling. After Grand et al. (1997).

For a color version of this figure, see plate section.

In summary, the three-dimensional images of the mantle derived from seismic tomography strongly support the hypothesis that at least some slabs penetrate through the transition zone and sink into the lower mantle. Although it is uncertain if every subduction zone has a lower mantle extension, there is now good evidence that large quantities of slab material are present in the lower mantle. What is still uncertain is whether some subducted slabs sink all the way to the core–mantle boundary or are assimilated into the lower mantle before reaching the CMB.

Question 3.12: *Do some subducted slabs sink all the way to*
the core–mantle boundary?

Figure 3.31 provides evidence that at least some slabs sink all the way to the CMB. The figure shows cross-sections of mantle P-wave (A) and S-wave (B) velocity
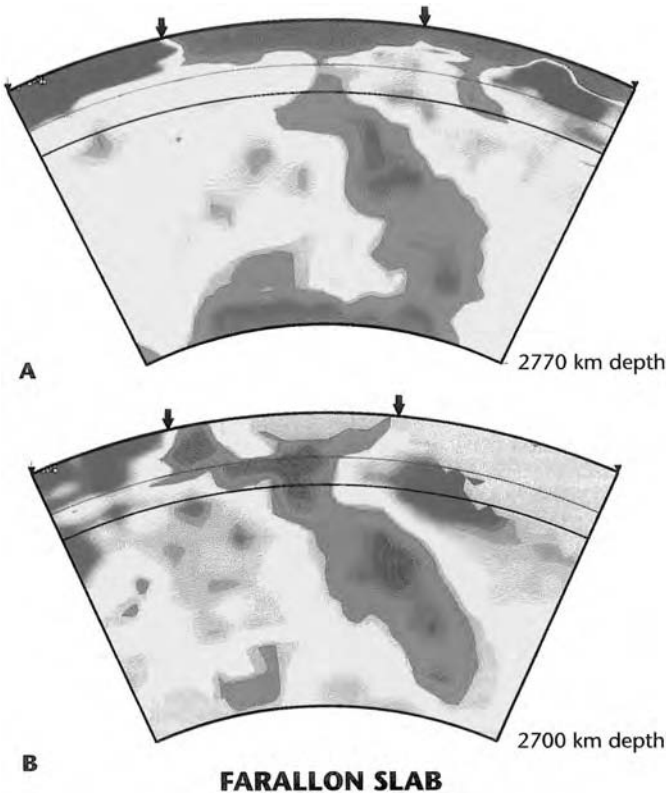

FARALLON SLAB

Figure 3.31. P- (A) and S- (B) wave velocity anomalies (relative to the global mean) at depths from the surface to the CMB in an approximately E–W cross-section of the mantle through the southern U.S. Reds indicate slow velocities, blues fast velocities. Differences in structure between the P-wave and S-wave models in the transition zone and near the CMB may be due to different data sampling in the two studies. The thin black lines are at depths of 410 and 660 km. After Grand et al. (1997).

For a color version of this figure, see plate section.

variations in an approximately east–west cross-section through the southern United States (Grand et al., 1997). The large slab-like fast seismic anomaly is probably the descending Farallon plate that subducted over the past approximately 100 Myr. The anomaly, and by implication the Farallon plate, extends all the way from the surface, through the depth of the 660 km seismic discontinuity and through the entire lower mantle to the CMB.

While there is evidence that at least some slabs sink all the way to the CMB, the shape of the slabs must change between about 1,800 km depth and the CMB since the long narrow structures seen in the tomographic images in the mid-mantle are replaced by broad large-scale features just above the CMB (Figure 3.30, van der Hilst et al., 1997). We will see in Chapter 10 that a change in the morphology of mantle downwellings from sheet-like to cylindrical with increasing depth is in accord with some numerical simulations of three-dimensional mantle convection (Bercovici et al., 1989a).

3.9.3 *Lower Mantle Seismic Heterogeneity*

Globally distributed seismic heterogeneity in the lower mantle was first investigated by Julian and Sengupta (1973) and by Dziewonski et al. (1977), both using P-wave travel time anomalies. The first global models of lower mantle structure were constructed using P-wave travel time residuals (Dziewonski, 1984). Although P waves continue to be used for lower mantle seismic tomography, there is an increasing number of lower mantle seismic tomography models based on S-wave travel times and waveforms. Su and Dziewonski (1997) have carried out simultaneous three-dimensional inversions of the seismic shear and bulk velocities in the mantle and find that the variations in these velocities are uncorrelated in the lower mantle.

Figure 3.32 shows a three-dimensional seismic shear wave velocity model of the mantle complete to spherical harmonic degree 16 by Masters et al. (1996) at different depths, contoured uniformly to permit comparison among the depths. Uniform contouring emphasizes the fact that large-scale mantle seismic heterogeneity is concentrated in the upper mantle above 300 km depth and near the base of the mantle, within 300 km of the CMB. Large-scale seismic heterogeneity is weakest from the base of the transition zone to about 2,200 km depth. This behavior is summarized in Figure 3.33, which shows the rms amplitude of the seismic heterogeneity from the model in Figure 3.32 as a function of depth. The shape of the profile of rms heterogeneity agrees qualitatively with the shape expected for whole mantle convection, with heterogeneity maxima in the surface and basal boundary layers separated by a broad heterogeneity minimum in the interior. The interior minimum corresponds to the region of the mantle where heterogeneity is reduced by the mixing action of convection. It is significant that the model does not exhibit an interior heterogeneity maximum at the boundary between the upper and the lower mantle, as would be expected if mantle convection were separated into two layers.

Although this and other similar tomographic models of the lower mantle contain structure that varies irregularly with depth, the largest-scale structure in these models is nearly independent of depth and is common to nearly all P-wave and S-wave models. As shown in the comparison of different models in Figure 3.34, the largest-scale structure in the lower mantle consists of a ring of high seismic velocity around the Pacific basin, separating two broad circular regions of lower than average seismic velocity, one centered in the south central Pacific and the other beneath Africa. The slow seismic anomaly centered beneath southern Africa is particularly well revealed in the S-wave model of Figure 3.31 where it extends from the base of the mantle to near 1,000 km depth. The high seismic velocity ring is
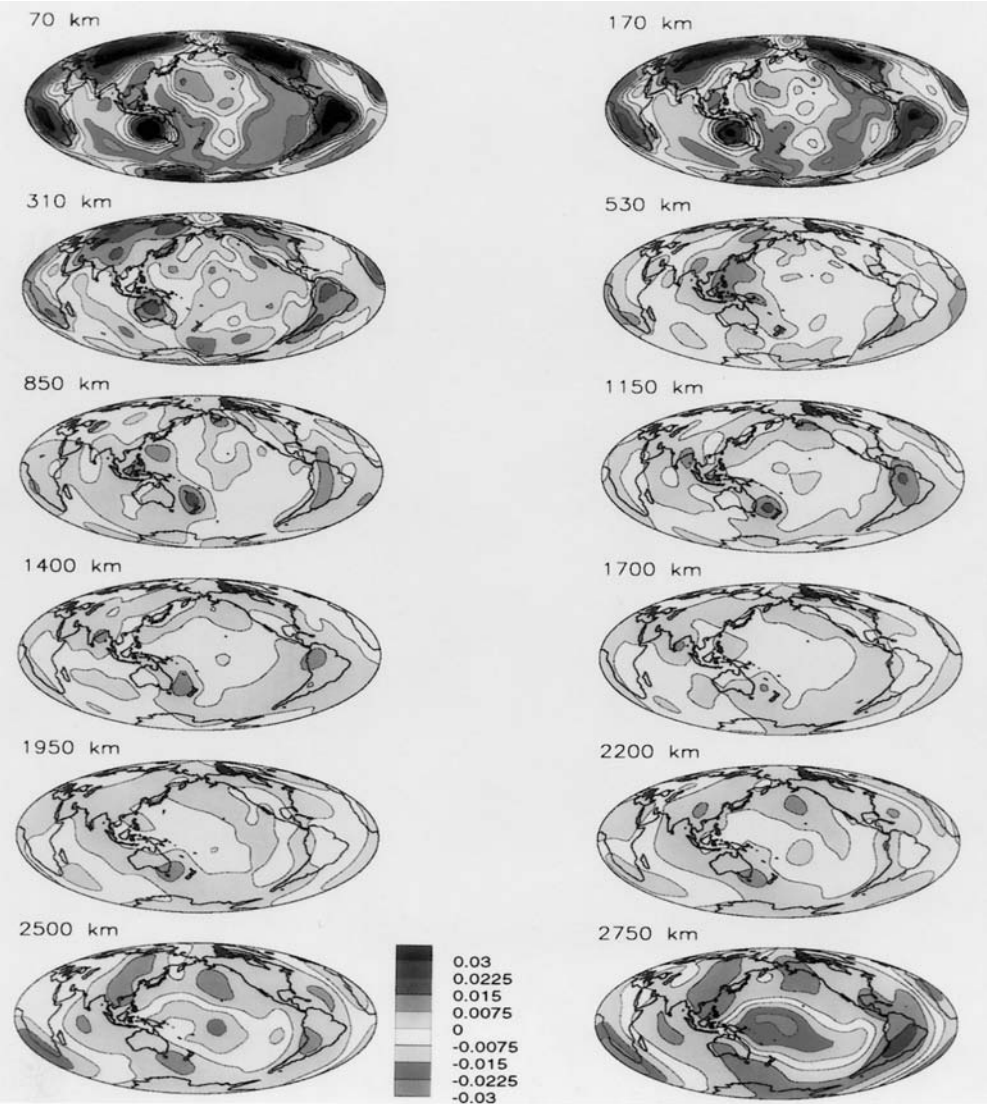
Figure 3.32. Perturbations in shear wave velocity V_S through the whole mantle at 12 depth slices, complete to spherical harmonic degree 16 determined by Masters et al. (1996).

For a color version of this figure, see plate section.

nearly a great circle, so the heterogeneity is approximately a spherical harmonic P_2^2 pattern. The ring of high seismic velocity corresponds roughly to the global system of Mesozoic and early Cenozoic subduction zones. If we interpret high seismic velocities as the result of low temperatures, then this structure indicates a ring of descending flow throughout the lower mantle, located approximately beneath long-lived convergent plate boundary regions. The broad columnar regions with lower than average seismic velocity beneath the Pacific and African plates then correspond to hotter than average mantle with upwelling velocity. The spherical harmonic degree 2 pattern of seismic velocities in the lower mantle, inferred

Figure 3.33. The perturbations in shear wave velocity V_S as a function of depth (solid curve) through the whole mantle from the tomography model of Masters et al. (1996) shown in Figure 3.32. The dashed curve is from the seismic tomography model of Su et al. (1994) for comparison.
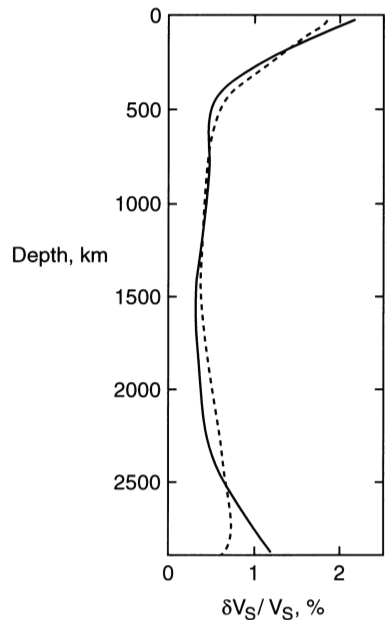
to be the major pattern of upwellings and downwellings (Figure 3.34), is one of the most important results from mantle seismic tomography.

In the global tomography models, the level of heterogeneity in the layer D″ is higher than elsewhere in the lower mantle, although the large-scale pattern in D″ is similar to that in the rest of the lower mantle, as shown in Figure 3.34. We have already discussed the evidence for a seismic discontinuity at the top of D″, several hundred kilometers above the CMB. Variations in the height of the discontinuity, relative to the pattern of large-scale seismic heterogeneity in the rest of the lower mantle, provide a clue to the nature and origin of D″.

The existing evidence indicates a positive correlation between D″ layer thickness, measured by the depth of the velocity increase, and lower mantle seismic velocity variations, as shown in Figure 3.35 from Kendall and Shearer (1994). The D″ layer is well developed around the margins of the Pacific where Figure 3.35 indicates that the lower mantle is anomalously fast. In the central Pacific, where global seismic tomographic images indicate slow lower mantle seismic velocities, the D″ layer discontinuity is not so well defined (Garnero et al., 1988). This correlation suggests that the discontinuity at the top of D″ is best developed beneath lower mantle downwellings. This is not the expected behavior if the D″ layer is an ancient, highly dense reservoir, because in that case, it should be best developed beneath upwellings. However, it is possible to explain this correlation if D″ consists of material accumulating at the base of downwellings in the lower mantle. Such a distribution implies continual influx of new material from above, and that D″ is an open system (Olson and Kincaid, 1991). The most likely source of dense material is former lithospheric slabs subducted into the lower mantle, which has obvious implications for the degree of mass transfer between upper and lower mantle. It is important to emphasize that the correlations on which these arguments are based are poorly established, and therefore any models based on them are highly speculative. A more secure interpretation requires further exploration of D″ and its spatial variations as well as improved resolution of lower mantle structure.
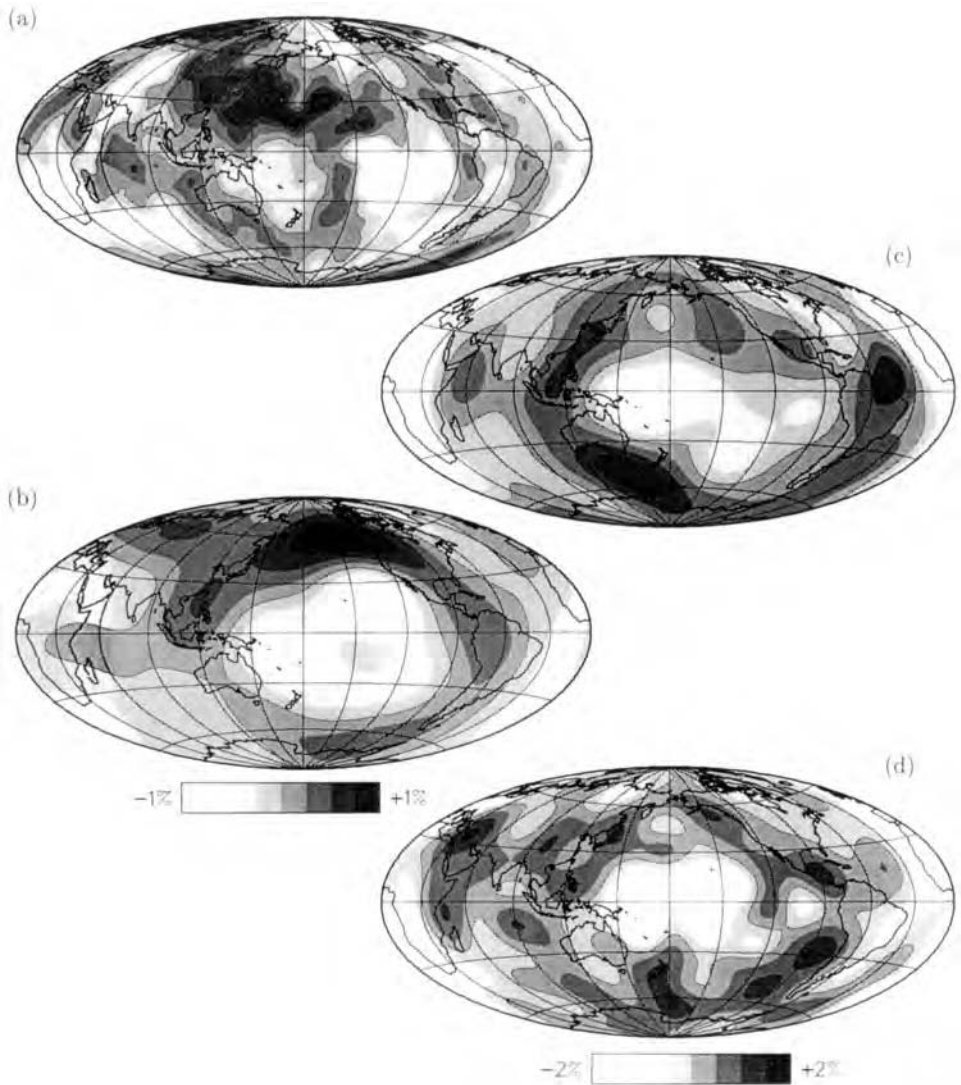
Figure 3.34. Comparison of four seismic tomography models of three-dimensional lower mantle structure. The ±1% scale applies to images (a) and (b); the ±2% scale applies to images (c) and (d). Image (a) is P-wave velocity variations at 2,566–2,900 km depth by Inoue et al. (1990); image (b) is S-wave velocity variations at 2,630–2,891 km depth by Tanimoto (1990); image (c) is S-wave velocity variations at 2,630–2,891 km depth by Masters et al. (1996); image (d) is S-wave velocity variations at 2,850 km depth by Su et al. (1994) for comparison.

3.9.4   Topography of the Core–Mantle Boundary

Question 3.13: *What is the topography of the core–mantle boundary?*

Another source of seismic heterogeneity that is directly coupled to the pattern of lower mantle convection is the shape of the core–mantle boundary. We have already mentioned the
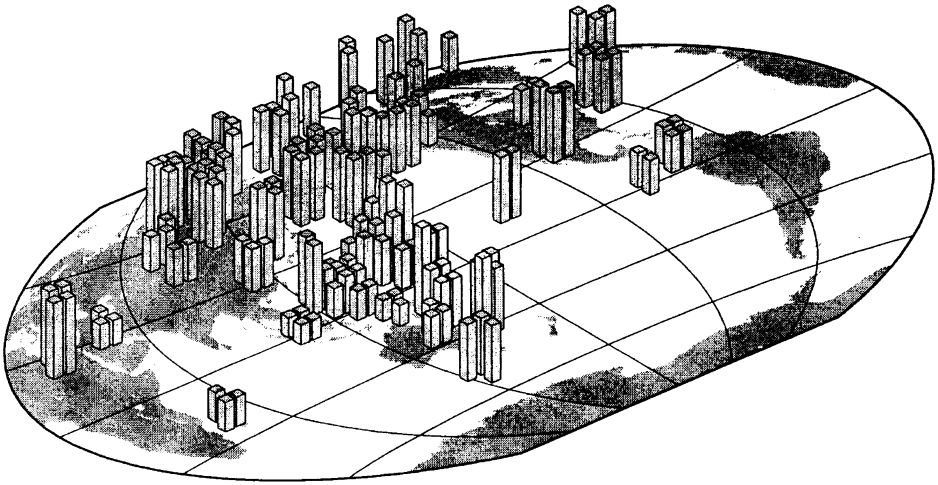
Figure 3.35. Lateral variations of the height of the top of the D″ layer compiled by Kendall and Shearer (1994).

observations of high-frequency P-wave scattering, from either the CMB, implying a rough CMB, or from small-scale heterogeneities distributed within D″.

Long-wavelength undulations of the CMB are an important constraint on the structure of flow in the lower mantle, since they are expected to be highly correlated with vertical motion. Unfortunately, long-wavelength CMB topography remains a controversial subject. Several investigators have used the travel times of transmitted and reflected core phases to map long-wavelength CMB heterogeneity, but there is little consistency among the results (see the review by Morelli, 1993). Using transmitted phases only, Creager and Jordan (1986b) interpreted travel time residuals as being due to either "inverse continents" – accumulations of heavy mantle material or "slag" – or light core material. However, the pattern of seismic heterogeneity they found is quite different from the results of Morelli and Dziewonski (1987), who used both reflections (PcP) and transmitted waves (PKP). In principle, a comparison of the travel time anomalies from these two phases can distinguish between CMB topography and heterogeneities in D″, because they are positively correlated for heterogeneities in D″ and negatively correlated for CMB topography. Furthermore, if the heterogeneity were due to core slag, then PcP should not show a travel time anomaly at all. Morelli and Dziewonski (1987) found negative correlation, indicating CMB topography as the source, and inferred a maximum range of ±6 km. This result has been questioned, because Earth rotation data constrain the P_2^0 and P_2^1 spherical harmonic components of the topography to be small (see below). An analysis of travel time anomalies by Doornbos and Hilton (1989), including bottom side CMB reflections (PKKP) in addition to the other phases, resulted in a model with relatively little CMB topography (1.5 km^2 variance, versus 20 km^2 from the previous studies), and with most of the heterogeneous structure located in D″.

The only firm evidence for anomalous topography on the CMB comes from observations of luni-solar nutations and solid earth tides (Gwinn et al., 1986) which indicate that the CMB departs from hydrostatic shape at spherical harmonic degree 2 by 0.5–0.9 km (Wahr and de Vries, 1989). Although the issue remains open, it seems that there may be only a small amount of CMB topography (Rodgers and Wahr, 1993; Vasco and Johnson, 1998), and the problem of delineating it is aggravated by lateral variations in the D″ layer.