I. GREAT EARTHQUAKES AND SEISMIC COUPLING

AT SUBDUCTION ZONES

II. THE STRUCTURE OF THE LOWERMOST MANTLE DETERMINED

BY SHORT PERIOD P-WAVE AMPLITUDES

Thesis by

Larry John Ruff

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

1982

(Submitted December 14, 1981)

ACKNOWLEDGEMENTS

During my stay at the Seismological Laboratory, I have had the benefit of a friendly atmosphere created by the faculty, staff, and students. I thank all of the people involved in this productive climate, and I hope that this atmosphere will continue for the benefit of all students at the Seismo. Lab.

My scientific training has been aided considerably by the close interaction with patient faculty members. In particular, I acknowledge the efforts of Hiroo Kanamori, Don Helmberger, and Don Anderson.

I would like to thank all of the people that have contributed to my growth, both personally and scientifically. It is truly difficult to select any names as my family and many friends have been important. However, Holly Eissler and Hiroo Kanamori have had the greatest influence in the above two respects the last two years. I must mention a special friend of the past six years, Eveline den Otter, who has given me love and support at times when others have found it more convenient to do otherwise. Thank you.

ABSTRACT

Part I

Seismic coupling has been used as a qualitative measure of the "interaction" between the two plates at subduction zones. Kanamori (1971) introduced seismic coupling after noting that the characteristic size of earthquakes varies systematically for the northern Pacific subduction zones. Great earthquakes ($M_W > 8.5$) occur in only a few subduction zones: notably the northern Pacific and South American subduction zones. A quantitative global comparison of many subduction zones reveals a strong correlation of earthquake size with two other subduction zone variables: age of the subducting lithosphere and convergence rate. The largest earthquakes occur in zones with young lithosphere and fast convergence rates, while zones with old lithosphere and slow rates are relatively aseismic for large earthquakes. Two other correlations are of interest; maximum depth of the continuous Benioff zone is correlated to lithosphere age, and horizontal length of the Benioff zone is correlated to convergence rate. The simplest explanation of these correlations is "preferred trajectory": the subducting slab descends into the mantle with the vertical and horizontal rates determined by the plate age and convergence rate respectively. The mechanism of preferred trajectory is also consistent with the obversation that back-arc spreading occurs behind subduction zones that are subducting old lithosphere at a slow rate.

The rupture process of a great earthquake indicates the distribution of weak and strong regions on the fault zone between the subducting and over-lying plates. The rupture process of three great earthquakes (1963 Kurile Islands, $M_W=8.5$; 1965 Rat Islands, $M_W=8.7$; 1964 Alaska, $M_W=9.2$) are studied by using WWSSN stations in the core shadow zone. The main result is that maximum earthquake size is determined by the asperity distribution on the fault plane (asperities are the strong regions that resist the motion between the two plates). The subduction zones with the largest earthquakes have very large asperities (the Alaskan earthquake is characterized by a giant asperity of length scale 150-200 km), while the zones with smaller earthquakes have small scattered asperities. This observation can be translated into a simple model of seismic coupling, where the horizontal compressive stress between the two plates is proportional to the ratio of the summed asperity area to the total area of the contact surface.

If asperity size determines earthquake size, and earthquake size is correlated to plate age and rate; then plate age and rate must be related to the asperity distribution. Plate age and rate can control asperity distribution directly by use of the horizontal compressive stress associated with the preferred trajectory. Indirect influences are many, including: oceanic plate topography and the amount of subducted sediments.

All subduction zones are apparently uncoupled below a depth of about 40 km, and the basalt to eclogite phase change in the down-going oceanic crust may be largely responsible. This phase change should

start at a depth of 30-35 km, and could at least partially uncouple the plates by superplastic deformation throughout the oceanic crust during the phase change.

Part II

The seismic velocities in the D" region (lowermost 200 km of the mantle) are recognized to be anomalously low, though the details of the velocity structure are not known. The details of D" are important, in particular whether a smooth velocity model is appropriate or not. A smooth decrease in the seismic velocities would be consistent with a thermal boundary layer at the base of the mantle. We have used the amplitudes of short period (T = 1 sec) P waves to investigate the internal structure of D". A short period amplitude data set is obtained by using underground nuclear events as sources and applying receiver corrections to the amplitudes. Receiver effects are largely responsible for the factor of ~ 8 scatter in the amplitudes of the North American WWSSN stations. Applying receiver corrections reduces the scatter to a factor of ~ 2, thereby providing a quantitatively useful amplitude profile into the core shadow. Using Soviet events and North American WWSSN statios, the D" layer beneath the north polar region is well sampled. The core shadow (at T = 1 sec) begins sharply at a distance of Δ = 95.5 and the slope of the amplitude decay is well defined. Also, the amplitudes decrease slightly from Δ ~ 87 to Δ ~ 90, then increase to Δ ~ 95. Synthetic seismograms are used to test various earth models,

with the important conclusion that the amplitudes from smooth D" models with a nearly constant velocity in D" decay too slowly in the shadow. This mismatch cannot be satisfactorily explained by random forward scattering or a thin low-Q layer within D". Anelastic calculations show that a thin low-Q layer in D" decreases the amplitudes gradually before the shadow, with little effect on the decay slope in the shadow. All of the features of the observed amplitude profile can be explained as the interference effects of a model that has a low velocity zone in the upper part of D" followed by a normal velocity gradient in the lower part of D". This type of model (POLAR series) also explains the scatter often observed in dT/dΔ beyond Δ ~ 90. The interference effects and required velocity changes in D" are small, and long period amplitudes will respond only to the averaged velocity gradient in D". The POLAR models imply a compositional and/or phase change at the top of D". Thus, the preferred seismological model does not allow the D" region to be interpreted as a single thermal boundary layer between the mantle and core.

TABLE OF CONTENTS

1

Part I

Great Earthquakes and Seismic Coupling

at Subduction Zones

Chapter 1

Introduction to Part I

Although the following three chapters form a reasonably coherent sequence, it is important to unite the focus of these studies. This general introduction provides the opportunity to consider the broad context. The following three questions are of particular interest: what is the scientific problem, what is the contribution of this thesis, and what are future areas of study?

The scientific problem is to explain the occurrence and role of great earthquakes at subduction zones. Needless to say, this thesis does not rigorously solve the above problem. However, the following chapters take a few steps in a direction that seems to be fruitful. Before describing these results, a short review of great earthquakes and seismic coupling is presented.

The earthquake catalogue of Gutenberg and Richter (1954) used a surface-wave magnitude to measure the size of earthquakes. The surface-wave magnitude scale is not a useful measure of earthquakes with a magnitude greater than 8, and the introduction and use of the seismic moment with the subsequent development of the moment magnitude, M_W (Kanamori,1977a), allows a more meaningful comparison of the largest earthquakes. Nearly all great earthquakes (M_W>8.5) occur in subduction zones, with focal mechanisms consistent with the underthrusting of the oceanic lithosphere. These great earthquakes are not evenly distributed amongst the subduction zones (see Figure 2.1). The largest events have occurred primarily in the North Pacific and South America, while other subduction zones, such as the Marianas, are relatively quiescent with no

major earthquakes. Kanamori(1971) noticed that the size of earthquakes around the northwestern Pacific seems to vary systematically from Alaska to the Marianas, and related this change in earthquake size to a change in "seismic coupling". Kanamori suggested that seismic coupling could be related to the properties of the contact surface between the subducted and overlying plates: in particular, the degradation of the contact surface. Later, Kelleher et al.(1974) surveyed the size of earthquakes at subduction zones, and related earthquake size to the width of the contact plane at shallow depth. They noted that the two largest recorded earthquakes, the 1960 Chile (M_W=9.5) and 1964 Alaska (M_W=9.2) earthquakes, occurred in regions of a distinct broadening of the plate contact zone. Expanding upon the previous work, Kanamori (1977b) related the variation in earthquake size from Alaska to Marianas to the amount of aseismic slip, and concluded that a smaller earthquake size indicated a larger proportion of aseimic slip. Thus, at least in the northwestern Pacific, "seismic coupling" can be interpreted as the proportion of seismic slip with a strong coupling indicating that the relative motion of the plates is accommodated mostly by earthquakes.

In a search for some explanation of the variation in earthquake size, the characteristic maximum earthquake sizes in subduction zones are quantitatively compared to various other physical variables of subduction zones (Chapter 2). The main result of this comparison is that the largest earthquakes occur in subduction zones which are subducting young lithosphere at a fast rate while subduction zones with

old lithosphere and slow rate are relatively aseismic. This correlation implies that great earthquakes do not occur randomly, but instead are related to some "global" properties of the subduction zone, thus providing a framework in which to further categorize the great earthquakes. This result does not indicate why the great earthquakes are larger.

We want to understand the variation of earthquake size in terms of the causal mechanical conditions. Once again, we need to step back and develop the context in which we can examine the mechanical conditions of fault zones. The important concept is that fault zones are, in general, spatially heterogeneous in strength. Fault heterogeneity has been recognized rather recently to be fundamentally important in explaining many different aspects of earthquakes: for example, multiple event rupture, foreshocks and aftershocks, seismicity patterns, and the variation of seismic to aseismic slip along fault zones. Thus, the framework in which we view the rupture of great earthquakes is that fault zones are composed of strong and weak regions, and that the distribution of these regions at least partially controls the character of seismicity in a particular subduction zone. How the weak and strong regions interact during an earthquake cycle is currently controversial. One model of interaction with great appeal is the asperity model. In this model, the weaker regions fail by aseismic slip or small earthquakes as the fault plane is loaded by plate motions. Failure of the weak regions gradually concentrates the stress at the largest and

strongest asperities. These asperities then fail as the largest earthquakes in that particular region.

With the concept of fault heterogeneity in mind, we can now approach the subject of the occurrence of great earthquakes. In particular, we want to consider whether the variation in earthquake size reflects a systematic change in the mechanical conditions of the fault zone. This problem can be investigated by looking at the rupture history of great earthquakes as the time variation of moment release translates into a spatial variation of moment, thereby indicating the degree of fault heterogeneity. The results presented in Chapter 3 show an increase in the dominant asperity size as earthquake size increases. This conclusion is largely based on the end member of the 1964 Alaskan earthquake. The entire northeastern segment of the Alaskan fault zone, a length scale of 200 km, can be considered as one huge strong asperity. This conclusion implies that the largest earthquakes are not "accidents", but are controlled by the presence of large asperities that cause great earthquakes. This result gives more confidence in the significance of the correlation between earthquake size and plate age and rate. (Note: The history of the Colombian subduction zone is somewhat problematic in this respect.)

In Chapter 4, we pursue the results of Chapters 2 and 3, which when combined imply that plate age and rate control asperity size. At this time, there is no physical quantitative model that relates plate properties to asperity size. We can only discuss several possible

connections that are consistent with the observed correlations. Also, a simple model of seismic coupling is presented in Chapter 4. This model connects earthquake size to the regional tectonic stress level via asperity length scale, and also offers an explanation of why the plates are seismically uncoupled below a depth of 40 km.

The contributions of this thesis can be summarized as follows: (i) there is a correlation between maximum characteristic earthquake size, convergence rate, and age of the subducting lithosphere, (ii) the size of large earthquakes in subduction zones is controlled by the asperity distribution in a simple way, i.e., a huge dominant asperity causes a giant earthquake. A coherent picture has thus emerged on seismic coupling at subduction zones.

What are the areas of interest for future work? It is important to remember that the conclusions of Chapter 3 are based on just a few earthquakes and can only be considered as a general trend. There are exceptions to the correlation between earthquake size, age and rate, and undoubtedly there will be exceptions to the pattern of larger asperity→ larger earthquake size. Exceptions might be understood as: a different interaction between adjacent large asperities, restrictions due to lateral boundaries, or from mechanisms not yet known. An obvious problem is understanding what asperities actually are, for example, geometric irregularities, variable coefficient of friction, or some other cause. The asperity model is a kinematic description. Perhaps a dynamical asperity model can be addressed in the future. This

development is necessary before earthquakes can be predicted on a deterministic basis. However, what is more significant at this point in time, and indeed exciting, is that the simple kinematic asperity model, utilizing the basic elements of weak and strong regions on the fault plane, explains quite a range of earthquake behavior; and serves to connect earthquake size, rupture history, recurrence times, and regional tectonic stress. This thesis hopefully contains a few important steps along the path toward a scientific understanding of subduction zone earthquakes.

Chapter 2

Seismicity and the Subduction Process

2.1. Introduction

Since the advent of the plate tectonics model, trench-island arc systems and seismic Benioff zones have been interpreted as features related to the subduction of oceanic lithosphere. Understanding the mechanics and dynamics of the subduction process in view of these observed features has proved to be difficult. As one approach to this problem, various models have been concerned with modeling the features of a composite subduction zone which combines the pertinent features. Another method is to compare the subduction zones of the world noting the similarities and differences in the essential features. We have employed the latter method and include the strength of coupling at plate boundaries as one of the essential features. We find that there is a correlation between Benioff zone geometry, convergence rate, age of the subducting slab, strength of coupling, and formation of marginal (back-arc) basins.

In a global view of subduction zones, two key discriminating physical features are the extent and geometry of the Benioff zone, and the presence or absence of a marginal basin (Uyeda, 1977). The Benioff zone geometry relates to the behavior of the subducted lithosphere within the mantle, while marginal basin formation presumably results from the interaction of the slab with the surrounding mantle and crust. Therefore, these features should be at least partially diagnostic of the mechanics of subduction. There is considerable variation in these two

features between subduction zones: the slab penetration depth varies from 100 to 700 km and the slab dip angle ranges from nearly horizontal to vertical. Also, marginal basin activity ranges from active formation to the absence of any back-arc spreading.

Studies attempting to relate these observed features to causal parameters have had some success. In particular, there appears to be a correlation between the maximum depth of the continuous Benioff zone and the age of the subducting oceanic lithosphere (Vlaar and Wortel 1976, Wortel and Vlaar 1978). An approximate relationship between the down-dip lengths of slabs and lithospheric convergence rates was found by Isacks et al. (1968). This relationship has been further elaborated by a proposed correlation between the total lengths of subducted slabs and the products of lithosphere ages and convergence rates (Molnar et al., 1979). Luyendyk (1970) investigated the variation in dips of down-going plates, and proposed a relation between dip and convergence velocity which assumed a constant vertical velocity for slabs. With regard to marginal basin formation, Kanamori (1977b) proposed that the Philippine Sea plate might have been formed by a complete decoupling of the slab from the upper plate, leading to a fast retreat of the trench line. (See also Wu, 1972; Uyeda and Kanamori, 1979.) Molnar and Atwater (1978) suggested that some relationship exists between marginal basins, convergence rates and oceanic lithosphere ages. The results of this chapter substantiate and combine the above suggestions.

Another key feature of subduction zones is their seismic character, indicated by the size of the largest, shallow, thrust earthquakes.

Kanamori (1971) noted the variation in this property for the northwestern Pacific , and related it to the nature of the mechanical coupling between the oceanic and continental plates and the formation of the Philippine Sea plate. Kelleher et al. (1974) reviewed the variation in subduction zone earthquake size, and proposed a correlation between earthquake size and the width of the contact zone. The world-wide variation in seismic character is indeed significant, from the South America and Alaska regions in which great earthquakes occur, to the Mariana and Izu-Bonin subduction zones which lack comparably large interplate thrust events. This contrast in regional seismic character, which represents more than two orders of magnitude in the characteristic energy release of subduction earthquakes, has not been fully appreciated. These differences are interpreted as representing significant variations in coupling between plates in subduction zones. Uyeda and Kanamori (1979) conducted a global survey of subduction zone features and recognized the importance of these variations in seismic coupling. In this paper, we have parameterized the coupling strength and have sought a quantitative relationship between the physical features of subduction zones.

2.2. Large earthquakes and strength of coupling

The relative size of earthquakes is typically described by a magnitude scale, with the 20-second surface wave magnitude, M_S, commonly used for the larger earthquakes. This magnitude scale is inadequate for

the largest earthquakes, however, as the scale saturates above $M_S=8$, thereby not truly representing the total energy release of large events. A more accurate indicator of the total energy release is the seismic moment $M_o = \mu A d$, where μ is the rigidity, A the fault area, and d the average displacement. In order to remedy the saturation problem of the M_S scale, Kanamori (1977a) devised a new magnitude scale for large events, denoted by M_W, in which the magnitude is determined by the seismic moment. This magnitude scale connects smoothly to the M_S scale at magnitude 8. As the moments of the largest instrumentally recorded earthquakes have been determined (see Kanamori, 1977a), the M_W scale allows us to compare the relative sizes of these large earthquakes. The locations and magnitudes of the largest events are shown in Figure 2.1.

One feature apparent in Figure 2.1 is the variation in the size of the largest event occurring within the various subduction regions, as pointed out by Uyeda and Kanamori (1979). For example, while South America is characterized by very large events, other regions such as the Marianas or the Scotia arc appear to be relatively quiescent. This scale of variation in seismic behavior will be included in our global view of the subduction process.

The characterization of subduction regions must be done with some reserve, due to the possibility that the largest earthquake in a particular region may not have been recorded, considering the limited period of observation. For subduction zones where the historical data of the past several hundred years are available (e.g., Japan, Kuriles, South America), the instrumental record from 1900 onward appears to
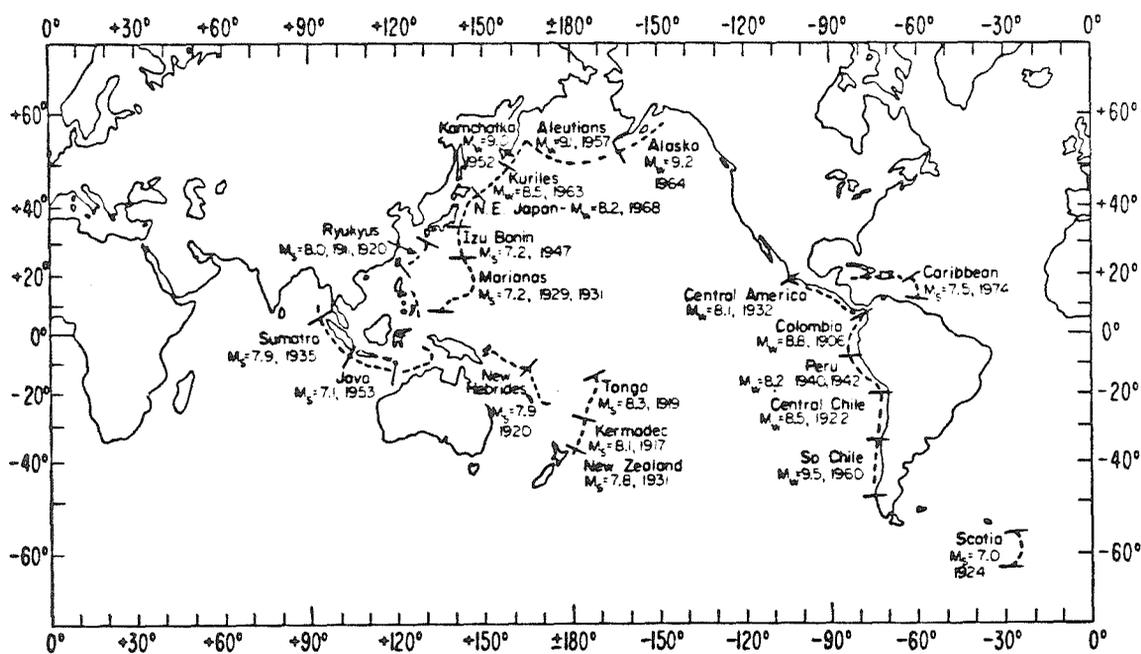
Figure 2.1. Subduction regions used in this study. The date and M_S or M_W value of the largest event within each subduction region are shown.

be representative of the seismic activity of the respective regions. That is, at least one typical large event has occurred since 1900 in each region. For subduction zones with an incomplete historical record and no occurrences of great earthquakes since 1900 (e.g., Izu-Bonin, Marianas, Java), the instrumental record may underestimate the seismic activity. However, historical tsunami data and an anomalously low frequency of earthquakes below magnitude 7 indicate that seismic activity is indeed low in these regions (see Kelleher and McCann, 1976). Though the particular M_W values that we assign to the subduction zones may be modified by an additional 100 years of observation, we conclude that the presently available values provide a basic representation of the varying seismic character of these zones.

The strength of coupling between upper and downgoing plates may be regarded as the product of the area of contact and the average stress on the contact zone. Since large thrust earthquakes along subduction zones represent a stress release on these contact zones, the seismic moment (and therefore M_W) indicates the strength of coupling. The provision of an explicit relationship between earthquake moment and strength of coupling would be difficult due to many complications, and the result would certainly be model-dependent. Consequently, we will use the observed variation in M_W as a representation of the variation in the strength of coupling.

Thus, we have taken the M_W of the largest earthquake occurring in a subduction region as a characteristic property of that region. An alternate measure would be to take the cumulative seismic moment (or its

logarithm) instead of the M_W of the largest earthquake. However, the contribution from the largest earthquake usually dominates the cumulative moment. In Appendix A, the cumulative seismic moment is calculated and corrected for subduction zone length along the trench axis. The effective M_W for a few subduction zones does change; however, the conclusions remain the same. For simplicity, we use the maximum M_W in the following discussion.

2.3. Data and correlation

We tested for correlations between the physical features of subduction zones by using linear multi-variable regression. We included the subduction zones for which we could reliably determine the parameters as listed in Table 2.1. The ages of subducting oceanic lithosphere are mostly from deep-sea drilling data and are consistent with values used elsewhere (Vlaar and Wortel, 1976; Molnar and Atwater, 1978). "Depth" refers to the maximum depth of the continuous Benioff zone. Hence, the deep and isolated zones of seismicity below South America and the Fiji Plateau are not considered to be related to the recently subducted lithosphere at the South America and New Hebrides trenches. "Length" is the distance from the trench line to the furthest lateral extent of the continuous Benioff zone.

The convergence rates are taken from Model AM1 (Minster et al., 1974). We have used the convergence rates determined by Seno (1977) for the Marianas and Ryukyus subduction zones. Also, we have used

TABLE 2.1

Subduction zones and parameters used in this study

Zone	Seismicity (M_W)	Depth (km)	Length (km)	Age (My)	Rate $(cm\ y^{-1})$
Marianas	7.2	700	300	150	4.0
Java	7.1	650	550	135	7.1
Izu-Bonin	7.2	550	500	150	6.1
N.E. Japan	8.2	600	1200	130	9.7
Tonga	8.3	650	600	120	8.9
Kermadec	8.1	570	400	120	6.4
Kuriles	8.5	625	800	100	9.3
Kamchatka	9.0	625	800	80	9.3
New Zealand	7.8	350	270	120	5.5
New Hebrides	7.9	270	170	60	2.7
Ryukyus	8.0	280	380	60	5.6
Aleutians	9.1	280	200	60	7.5
Sumatra	7.9	200	400	80	6.6
Alaska	9.2	140	450	40	5.9
Central America	8.1	200	200	45	8.0
Central Chile	8.5	250	550	50	11.0
S. Chile	9.5	160	500	20	11.1
Peru	8.2	200	700	45	10.0
Caribbean	7.5	250	280	100	2.0
Scotia arc	7.0	180	200	65	2.0
Colombia	8.8	150	220	20	7.7

a convergence rate for the New Hebrides that assumes the Fiji Plateau is decoupled from the Pacific plate (Isacks et al., 1969). This results in a rather large reduction of the New Hebrides convergence velocity.

We have included as many subduction zones as possible since we want to see the full range of subduction behavior. The Philippines and New Britain-Solomon Islands regions are excluded from this analysis. These two regions are quite complex and involve the interaction of more than two plates, possibly combined with subduction polarity changes. Other zones not considered are the Andaman Sea region and the Mediterranean deep seismic zone, as these tectonic settings are rather unique. The former includes highly oblique subduction and unusual back-arc spreading (Eguchi et al., 1979), and the latter a continent-continent subduction zone.

The correlation coefficients, r_{ij}, indicate the correlation between any two variables. The correlation between three or more variables can be tested with multivariate regression. The correlation is given by the multiple correlation coefficient, $r_{1,mn...}$ where 1 is the dependent variable with m, n, and any additional variables treated as independent variables. The correlation coefficients for the variables listed in Table 2.1 are shown in Table 2.2. For the number of data points used in the analysis, significance at the 99% level corresponds to the absolute value of $r_{ij} = 0.549$ (i.e. the probability is 1% that the variables i and j are independent). Pairs of variables that correlate above this level are: age and depth, rate and length, seismicity and age, and seismicity and rate. The correlation coefficients for seismicity

TABLE 2.2

Absolute values of the correlation coefficients using the data in
Table 2.1. The variables are S, seismicity, D, depth, L, length, A
age, and R, rate. The coefficients are symmetric about the diagonal.

	S	D	L	A	R
S		0.287	0.209	0.627	0.629
D	0.287		0.505	0.837	0.118
L	0.209	0.505		0.287	0.631
A	0.627	0.837	0.287		0.229
R	0.629	0.118	0.631	0.229	

TABLE 2.3

The multiple correlation coefficients with seismicity as the dependent
variable, and combinations of the other variables as the independent
variables. The abbreviations are the same as Table 2.2. Significance
at the 99% level corresponds to a value of 0.633.

$r_{S,AR}$	$r_{S,AD}$	$r_{S,AL}$	$r_{S,RD}$	$r_{S,RL}$	$r_{S,DL}$
0.802	0.763	0.743	0.727	0.675	0.500

and rate, and seismicity and age are nearly equal, indicating that the combination of rate and age would better explain the variation in seismicity. This is verified by the multiple correlation coefficient in Table 2.3. By using both age and rate, the correlation with seismicity is improved from 0.630 to 0.802. There are other high values of the multiple correlation coefficient in Table 2.3; however, these coefficients include either age or rate as one of the independent variables. The regression analysis shows that age and rate are the only sigificant variables in explaining the variation in seismicity. The correlation between these three variables is shown in Figure 2.2, where the subduction zones are plotted with respect to the independent variables, age and rate. The correlation is indicated by the best fit regression plane. The general trend is for low seismicity to correlate with the combination of older oceanic lithosphere and smaller convergence rates, while younger lithosphere and larger convergence rates are associated with great earthquakes.

The most notable exceptions to this trend are the M_W values for Peru and Central America. As shown in Appendix A, the moment sum M_W is somewhat larger than the single event M_W for these two regions. Also, the South American points which plot in the upper right-hand corner of Figure 2.2 are not quite on the trend of the other points. Figure 2.3 plots the results when the Central and South American points are removed from the regression analysis. It is interesting to note that the Central and South American subduction zones are east-dipping. The only other east-dipping zones are the New Hebrides and Sumatra.
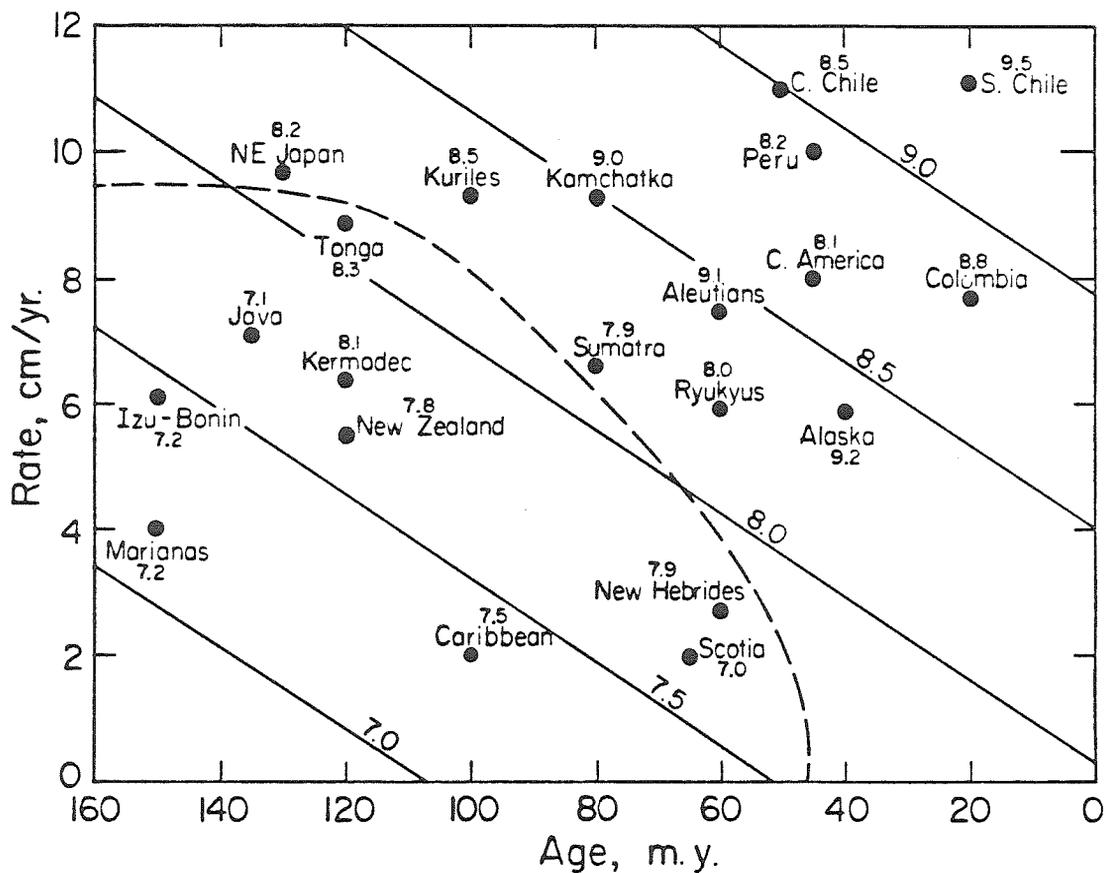
Figure 2.2. The relationship of seismicity to the two variables: convergence rate and age of the subducting oceanic lithosphere. The number at each subduction zone is the associated M_W, and the contours of constant M_W define the resultant plane from the regression analysis. The broken line in the lower left corner delimits the subduction regions where there is either confirmed or suspected back-arc spreading.
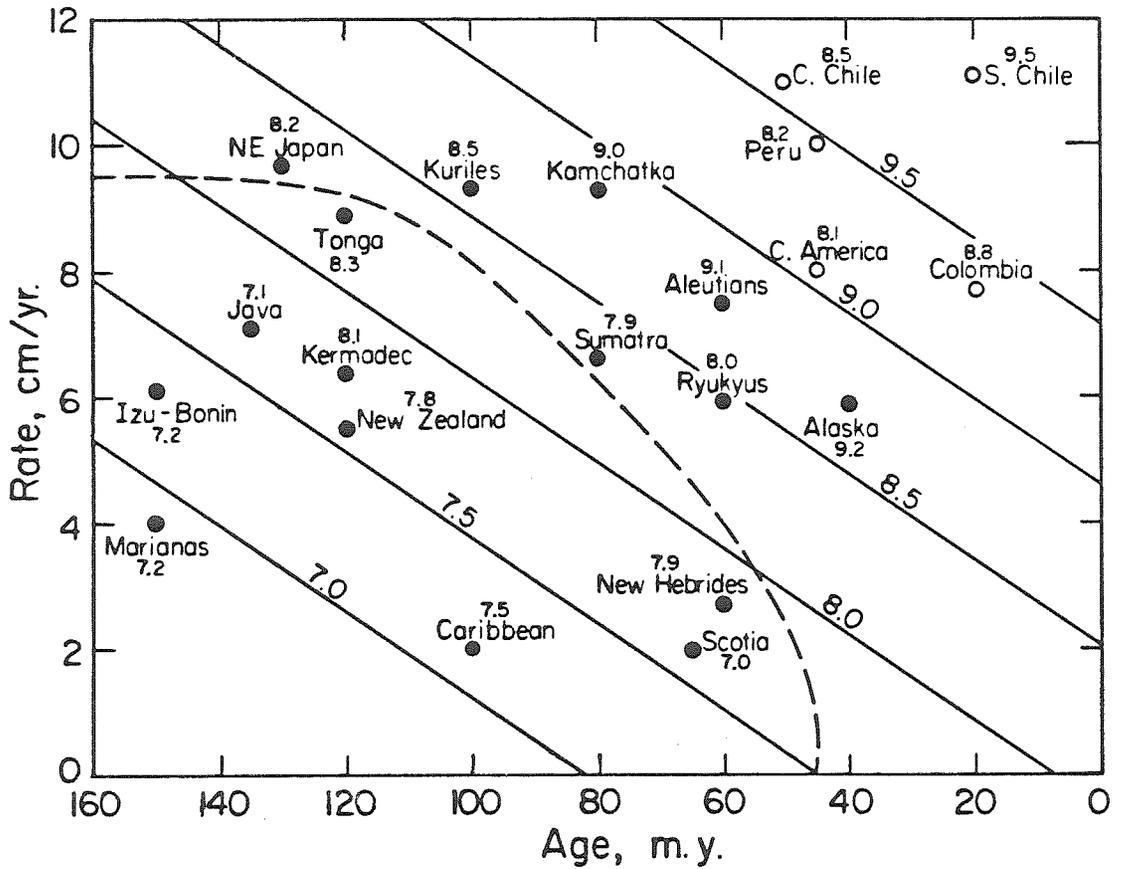
Figure 2.3. As Figure 2.2 except that the South and Central American points are not included in the regression analysis. The remaining 16 subduction zones conform quite well to the regression plane.

This apparent bias between east and west dipping zones was attributed to a global mantle flow by Uyeda and Kanamori (1979), but other explanations are certainly admittable.

There are other significant correlations within the data set. Depth of penetration is strongly correlated with lithospheric age (see Table 2.2). This correlation can be seen in Figure 2.4. Applying multivariate regression, rate enters the correlation at the 95% significance level. The multiple correlation coefficient with depth as the dependent variable and age and rate as the independent variables is $r_{D,AR} = 0.895$. This value is not appreciably larger than the simple correlation coefficient between depth and age, i.e. $r_{DA} = 0.837$. Therefore, there is only a weak dependence of depth upon rate. For completeness, this dependence is plotted in Figure 2.4 for different values of the convergence rate. The strong correlation between depth and age agrees with the results of Vlaar and Wortel (1976).

The other significant correlation in Table 2.2 is horizontal length and convergence rate, and these two variables are plotted in Figure 2.5. Using multivariate regression with length as the dependent variable, age enters the correlation at the 95% significance level, with $r_{L,RA} = 0.766$. It appears that this dependence is largely due to the outlying point of northeast Japan. There are no other significant (at 95%) multivariate correlations amongst the variables.

Figures 2.4 and 2.5 suggest that the penetration depth and horizontal length essentially depend on single variables, the lithospheric age and convergence rate respectively.
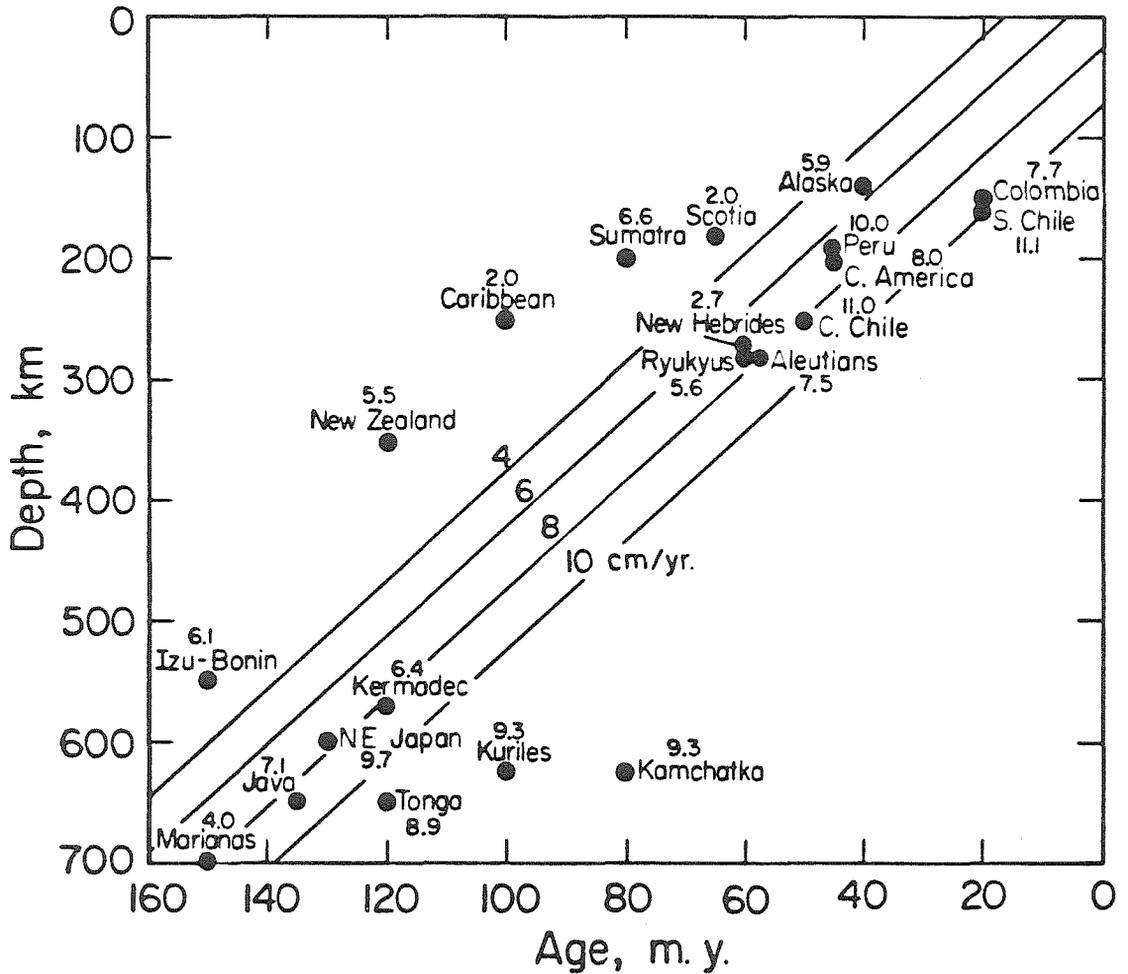
Figure 2.4. Depth versus age plot for the different subduction zones. (Treating the penetration depth of the Benioff zones as the independent parameter, the lithospheric age is the most significant variable.) Parallel lines show the regression solution for various values of the convergence rate. These contours do not appear to substantially explain the data, thus the convergence rate should be considered as an accessory correlation. Though there is a significant linear trend between depth and age, a bimodal distribution of penetration depths is also a good characterization of the data.

25
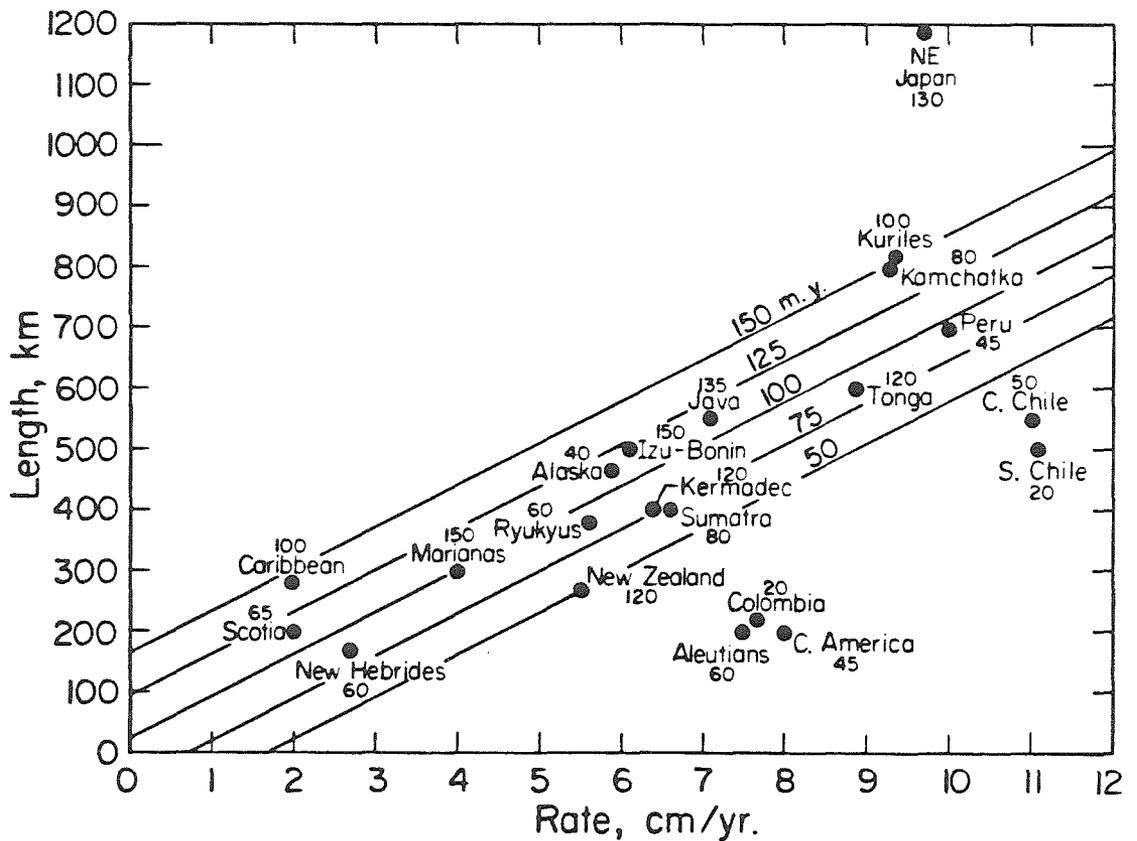
Figure 2.5. Plots of the subduction zone lengths versus convergence rates. (With the horizontal length of the slab as the independent variable, the convergence rate is the most significant variable, with the lithosphere age having secondary importance.) The linear correlation is given by the parallel lines, different lines corresponding to different values of lithosphere age (My).

On the other hand, Figure 2.2 indicates that two variables, age and rate, are required to explain the variation in the coupling strength. This is verified quantitatively by the simple and multiple correlation coefficients.

2.4. Discussion

We have sought the simplest explanation relating the strength of coupling to the properties of subducting slabs, though recognizing that other interpretations are possible. The fact that both coupling strength and Benioff zone horizontal and vertical extents depend upon two parameters, lithosphere age and convergence rate, suggests that there is a common mechanism affecting both Benioff zone geometry and coupling. Vlaar and Wortel (1976) explained the correlation between lithospheric age and penetration depth by the higher density of older slabs, which should therefore penetrate further into the mantle when subducted. This explanation seems reasonable if the intrinsic density of a subducted slab contributes to its downward penetration. As various types of evidence support this idea (e.g., McKenzie, 1969), it appears reasonable that initial density differences could affect both the rate of sinking and the level at which thermal assimilation occurs (Vlaar and Wortel, 1976). In view of this, the correlation of convergence rate and horizontal extent is related to the dependence of the horizontal velocity of a subducted slab on the convergence rate at the trench.

The preferred trajectory of subducting lithosphere in the mantle is

then determined by two factors, lithospheric age and convergence velocity. The horizontal and vertical rates should directly affect the dip of a subducting plate, with the total length of the slab dependent on the time scale of assimilation. As proposed by Molnar et al. (1979), this time scale may be mostly determined by the product of lithospheric age and convergence rate, hence a coupling between the age and rate upon the length and depth of the Benioff zone. The resolution of the slab's trajectory into horizontal and vertical rates is depicted in Figure 2.6. Deviations from the preferred trajectory can be caused by other factors which influence the geometry, such as global mantle flow (see Hager and O'Connell, 1978).

There is an apparent bimodal distribution in the penetration depths in Figure 2.4. The absence of slabs terminating between 300 and 400 km was attributed to difficulty in penetrating the phase change (presumably the 450 km discontinuity) by Vlaar and Wortel (1976). Another possibility, suggested by the relative aseismicity of this depth range in slabs which extend to below 600 km (Abe and Kanamori, 1979), is that once slabs penetrate below 300 km they can rapidly descend to 600 km (Anderson, 1979). We defer a detailed discussion of this distribution, since for our present purposes we are concerned only with the general trends.

The correlation of seismicity with convergence rate and lithospheric age implies a relationship between the strength of coupling and either slab geometry, or lithospheric age and convergence rate in some other manner. As a simple starting model, we consider
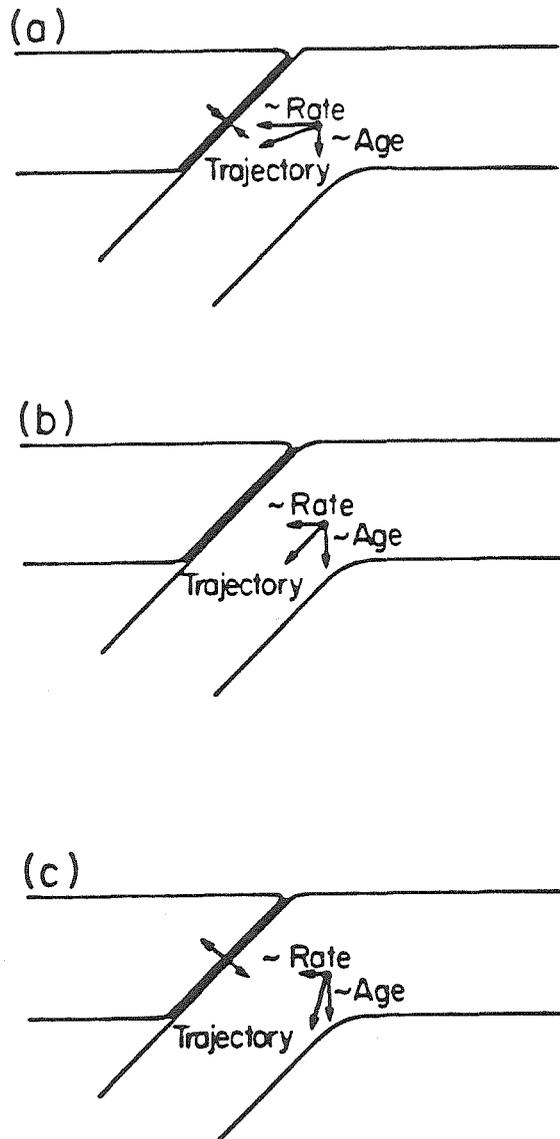
Figure 2.6. Schematic representation of how particular combinations of lithospheric ages and convergence rates might cause subducting slabs to have preferred trajectories, thereby affecting Benioff zone geometry. The preferred instantaneous trajectory of a slab is most certainly affected by other influences, such as the global mantle flow.

the strength of coupling to be dependent upon three factors: stress normal to the fault plane, the coefficient of friction, and the area of contact. There may be differences in the fault plane normal stress due to changes in the applied tectonic stress as slab geometry changes (Figure 2.6). Possibly the coefficient of friction changes systematically with convergence rate and age. Also, the area of coupling could change due to either a reduction in the cross-section as the slab dip angle becomes more vertical, or by a degradation of the upper plate's fault surface (Kanamori, 1971). Perhaps these different effects upon coupling strength operate in some combination, with one mechanism being dominant in certain subduction zones. In the next section on marginal basins it is suggested that the behavior of the downgoing plate could be the most important effect on coupling.

Seismic moment and coupling strength

As a brief departure from the main topic, the relationship between moment and coupling strength will now be considered in more detail. One might expect large earthquakes to have a higher stress drop, in addition to a larger fault area and displacement. The stress drops of earthquakes appear to be approximately equal over a large range of magnitudes (see, for example, Kanamori and Anderson, 1975). This constancy of stress drop can be qualitatively explained by an asperity fault model (Figure 2.7).

Instead of smooth planar fault surfaces, the asperity assumes
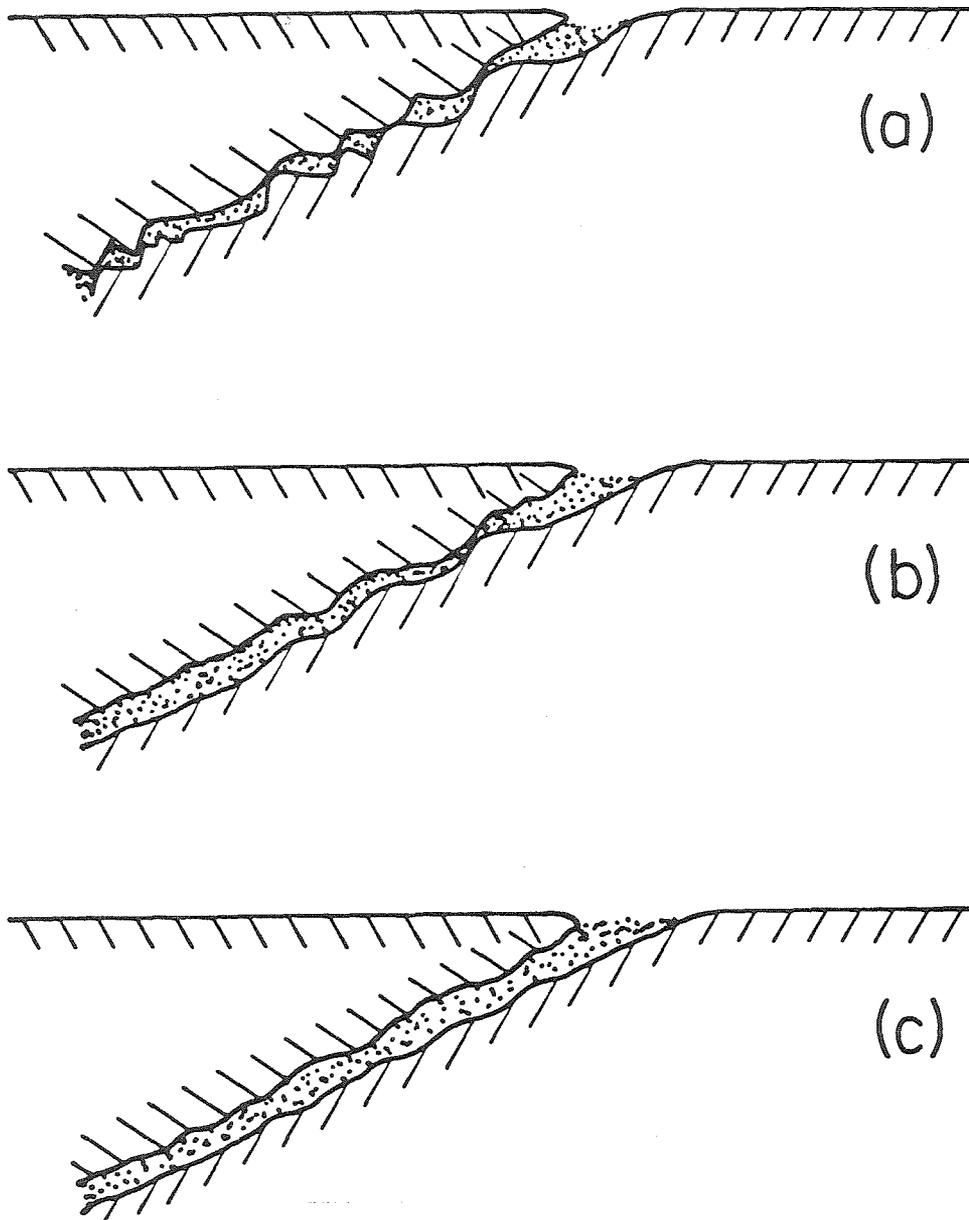
Figure 2.7. Schematic representation of three different asperity distributions. The dotted region between the upper and lower fault surfaces represents a low strength material. The tectonic stress is mostly distributed over the total area of asperity contact, with (a) representing a large area of asperity contact, thereby strong coupling. In (b) the tectonic stress is concentrated at one asperity and therefore the tectonic stress level achieved before a seismic event would be lower than in case (a). In (c), the tectonic stress is transmitted through the low-strength material. In this case the tectonic stress level will be the lowest, and the relative displacement will be mostly aseismic.

that most of the coupling is confined to the contact between
irregularities of the fault surfaces. The tectonic stress is
communicated across a smaller area than the total fault area, and so the
stress at the asperities is in general higher than the tectonic stress.
If there is a uniform characteristic rock strength, then the breaking
stress will be approximately equal. A larger asperity area increases
the contact area so that a higher tectonic stress is required to exceed
the asperity breaking stress. (Note: This model is discussed further
in Chapter 4.) Figure 2.7(c) shows that an aseismic weakly coupled
configuration results as a limit to a smaller asperity contact area.

Marginal sea formation

The consequences of the partial decoupling of plates as depicted in
Figure 2.6(c) could be marginal sea formation. If the preferred
trajectory of a slab becomes steep enough, there would be a tendency for
the oceanic plate to separate from the overlying plate. If this occurs,
the trench line will migrate oceanward; material must then be removed
from in front of the sinking plate and be supplied behind the moving
trench.

The details of the oceanic-plate bending process are not known and
the contribution of the upper plate cannot be fully assessed. However,
as the assumed cause of the plate bending is the gravitational sinking
of the slab, the plate should still bend if it is uncoupled from the
upper plate. Removal of material in front of the slab does not appear

to present serious difficulties. With regard to the supply of material behind the slab, kinematic models of the subduction process indicate that there is an induced corner flow due to viscous coupling of the oceanic plate with the surrounding mantle (e.g., Sleep and Toksoz, 1971; Andrews and Sleep, 1974; Toksoz and Hsui, 1978). This result precludes the possibility of material flowing updip along the slab. However, horizontal transport seems capable of supplying material to the corner region, as illustrated in Figure 2.8. Therefore, it appears reasonable that a weakly coupled contact zone could result in marginal basin formation.

We would therefore anticipate that low seismic coupling should be correlated with active marginal basin formation. the broken line in Figure 2.2 delimits the subduction zones in which active marginal basin formation is either known or suspected (see Karig, 1971; Uyeda and Kanamori, 1979, for reviews). The region of marginal basin activity is in the corner of lower convergence rates and older lithosphere, where coupling is weak in our model. We consider this result supportive of our seismic coupling model. It also emphasizes the fact that marginal basins are not currently forming behind subduction zones with great earthquakes. It should be noted that, even when the trench line is retreating, there is still relative motion across the contact zone. Therefore, there can still be occasional shallow thrust earthquakes, though not of great size. This supports the view of Kanamori (1977b) that aseismic subduction is the cause of diminished seismicity in certain segions, as opposed to suggestions that subduction
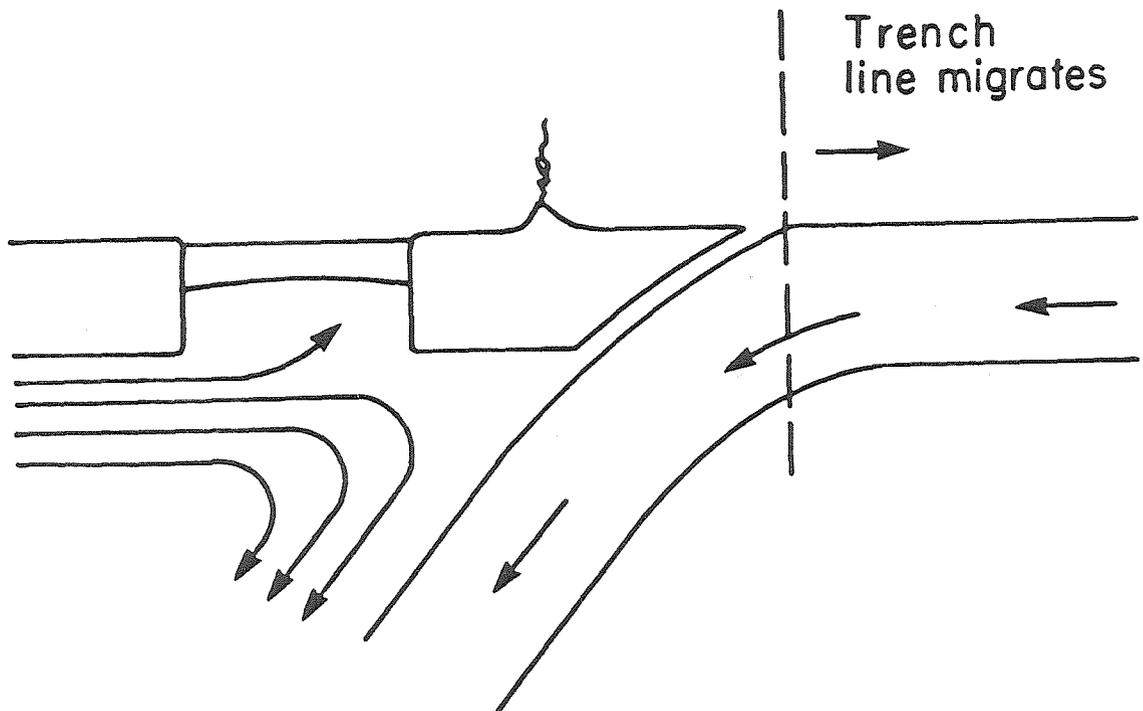
Figure 2.8. Back-arc spreading as a passive process. If the subducting slab bends and sinks such that there is a tendency for the trench line to migrate oceanward, then material must be supplied behind the volcanic arc. The proposed corner flow could possibly transport the necessary material to the island arc. Other studies have proposed that the corner flow itself forces the island arc to push the trenchline oceanward.

has ceased due to the presence of buoyant "gravity ridges" on the oceanic lithosphere.

Thus, in our view, marginal basin formation is inherently a passive process (Elasser, 1971; Molnar and Atwater, 1978). This result is in contrast to models of back-arc spreading derived from considerations of the induced corner flow (e.g., Toksoz and Hsui, 1978), in which back-arc spreading implies horizontal compression in the upper plate.

2.5. Conclusions

We have introduced strength of coupling as measured by M_W as an important physical feature of subduction zones, and find that both Benioff zone geometry and strength of coupling are significantly correlated to two variables: age of subducting oceanic lithosphere and plate convergence rate. The dependence of Benioff zone geometry can be explained by a preferred slab trajectory, with the horizontal and vertical rates determined by the convergence rate and age of the subducting plate. The strength of coupling can be affected by lithospheric age and convergence rate by either reducing the effective contact area between the plates, or perhaps by the preferred trajectory which transmits a small tectonic stress to the contact zone between the plates which can either enhance or detract the plate coupling. In the limit of weak coupling, aseismic subduction and marginal sea formation will take place. Other features of subduction zones, such as differences in chemical composition, might be better understood in the

context of the important physical features.

Chapter 3

The Rupture Process of Great Earthquakes

3.1. Introduction

The distribution and description of the largest earthquakes are of
fundamental interest in seismology. With the improvement in measuring
earthquake size by use of the seismic moment, the great earthquakes of
this century are now identified (see Table 3.1). The more recent
events have been studied using very long-period surface waves to
determine the seismic moment and fault orientation. Given the interest
in these large events, it is somewhat surprising that there has been no
systematic study of the rupture processes. The primary reason for the
lack of source time function investigations is that the P and S waves
are off-scale for the largest earthquakes. In this chapter, diffracted
P waves are used to recover the source time functions of the three most
recent great earthquakes (1963 Kurile Islands, 1964 Alaska, 1965 Rat
Islands) .

The source time functions have been determined for many smaller
earthquakes (i.e. M_W less than 8), and some of these earthquakes
consist of distinct multiple events which are usually interpreted as the
sequential breaking of asperities or barriers (regions of higher
strength on the fault plane) . It has commonly been assumed that great
earthquakes consist of a sequence of smaller (i.e. 'magnitude 8')
events, and the characteristic asperity size is the same as that for an
individual magnitude 8 event. In this view, a great earthquake occurs
when adjacent asperities along the subduction zone are all stressed
close to the failure stress, such that the failure of one asperity can

TABLE 3.1

The ten largest earthquakes in this century

(moments from Kanamori, 1977a)

Earthquake (year,location)	M_o ($\times 10^{27}$dyne-cm)	M_W
1960 Chile	2000	9.5
1964 Alaska	820	9.2
1957 Aleutian Islands	585	9.1
1952 Kamchatka	350	9.0
1906 Colombia-Ecuador	204	8.8
1965 Aleutian Islands	125	8.7
1950 Assam	100	8.6
1938 Banda Sea	70	8.5
1922 Chile	69	8.5
1963 Kurile Islands	67	8.5

trigger the adjacent asperities.

An alternative view is that great earthquakes occur where the characteristic asperity size is unusually large, and the failure of this asperity results in a substantially increased loading of the surrounding asperities, large or small, continuing the rupture over a larger area. Consequently, a great earthquake is directly associated with a larger asperity size. Hence, great earthquakes occur in places where the mechanical coupling between the two plates is stronger. Various approaches to seismic coupling and its relation to large earthquakes have been discussed and developed in: Kanamori (1971), Kelleher et al. (1974), Kelleher and McCann (1976), Kanamori (1977b), Toksoz and Hsui (1978), Uyeda and Kanamori (1979), Kanamori (1981), and Lay and Kanamori (1981).

Thus, in addition to the intrinsic interest in the rupture history of great earthquakes, the interpretation of the source time functions will test the asperity model of seismic coupling at subduction zones.

3.2. Diffracted P Waves

The rupture duration of large earthquakes ranges from 10 sec (magnitude 7.5 events) to possibly 300 sec for the 1964 Alaska and 1960 Chile events. Long period body waves are best suited for studying the rupture history as they are recorded in the period range from 10 sec to ~200 sec. The WWSSN network was established in the early 1960's to provide world wide coverage with standard instruments. This data base

has allowed the routine use of synthetic seismograms to model earthquake sources. The earthquakes are usually modelled by using seismograms recorded in the distance range Δ=30 to 90 degrees, where there are no complications in the P and S waves due to earth structure. In this distance range the elastic earth response can be adequately modelled by using geometric ray theory, i.e. there is no distortion of the waveshape. At distances less than 30 deg, the upper mantle triplications cause multiple arrivals in the seismograms. Upper mantle structure must be properly characterized before these recordings can be used for source studies. The core shadow zone starts at ~85 deg for long period P waves, and the P waves recorded beyond this distance have been diffracted around the core which causes a frequency dependent amplitude decay. The PKP and PKIKP phases which are recorded from Δ=120 to 180 deg are also not used for source studies as: there are two or more arrivals over much of the distance range, the amplitudes of these phases are fairly sensitive to structure at the core-mantle and inner core-outer core boundaries, and there is a rapid spatial variation in the PKP amplitude away from the PKP caustic at Δ~140 deg.

The long period P waves of the three great earthquakes are off-scale at virtually all of the WWSSN stations at distances less than 90 deg. Also, all PKP and PKIKP recordings are off-scale. The source history might be obtained from special well-calibrated low-gain long period instruments. However, for these three events we have been able to collect only one good record, a low-gain high-paper-speed strain record of the 1964 Alaska event recorded at Pasadena. Given this lack

of data, it is understandable why there have been no body wave source studies of these events.

We can exploit the fact that diffraction around the core attenuates the P waves, thereby providing on-scale long period recordings at WWSSN stations sufficiently far into the shadow zone. There are in fact two advantages in using diffracted P waves instead of the mid-mantle P waves. As seen in Figure 3.1, diffracted P waves have the longest time "window" in which to view the source time function. That is, at $\Delta = 110$ deg, there are four minutes in which no other major seismic phase arrives. As we expect source durations to be greater than a minute, it is convenient not to have later arrivals interfering with the source time function. The other advantage of diffracted P waves is that, since the higher frequencies are preferentially attenuated, diffraction acts as a natural smoothing filter. This is useful as we are interested in periods greater than 5 sec.

There are recordings of long period diffracted P waves for the three great earthquakes, and considering the availability and advantages of using diffracted P waves, an obvious question is why diffracted P waves have not been used previously. It would seem that the answer to this is "fear of the shadow zone". Just as seismograms at ranges less then 30 deg have been avoided due to structural complexities, diffracted P waves have been avoided as it has been assumed that reliable estimates of the source could not be obtained due to the complexity of the diffracted response. While the upper mantle response is unquestionably complicated, the diffracted response is not complex for long periods.
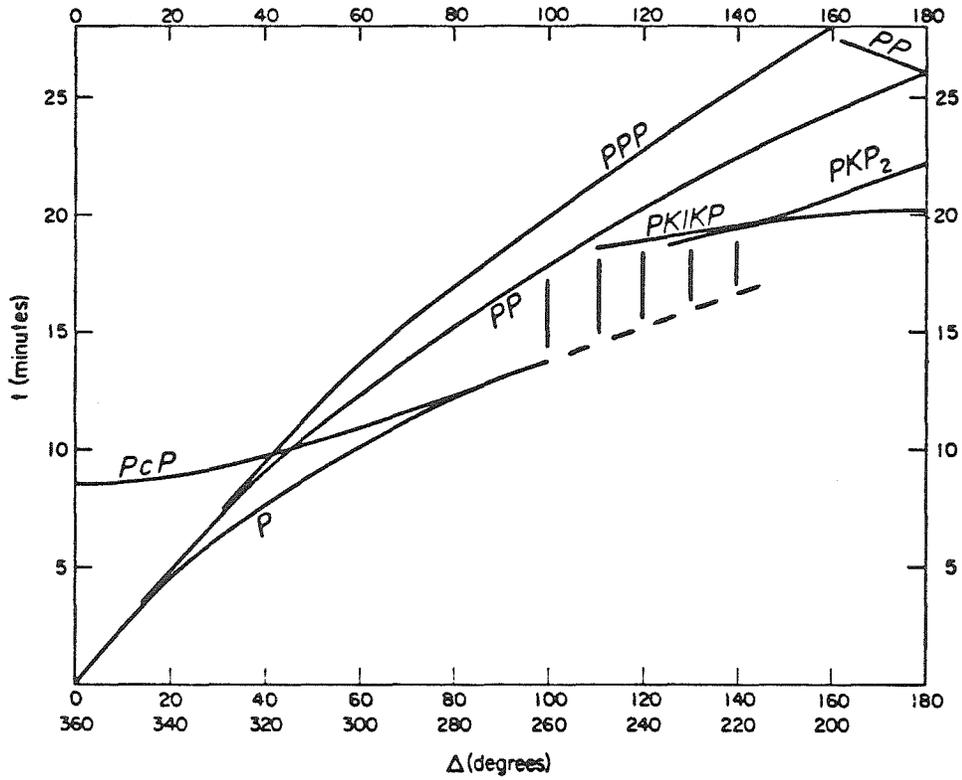
Figure 3.1. Simplified travel-time curves for P phases. Diffracted P is indicated by the dashed extension of P and PcP. There is a time window of 3 to 4 minutes after diffracted P in which no major seismic phase arrives.
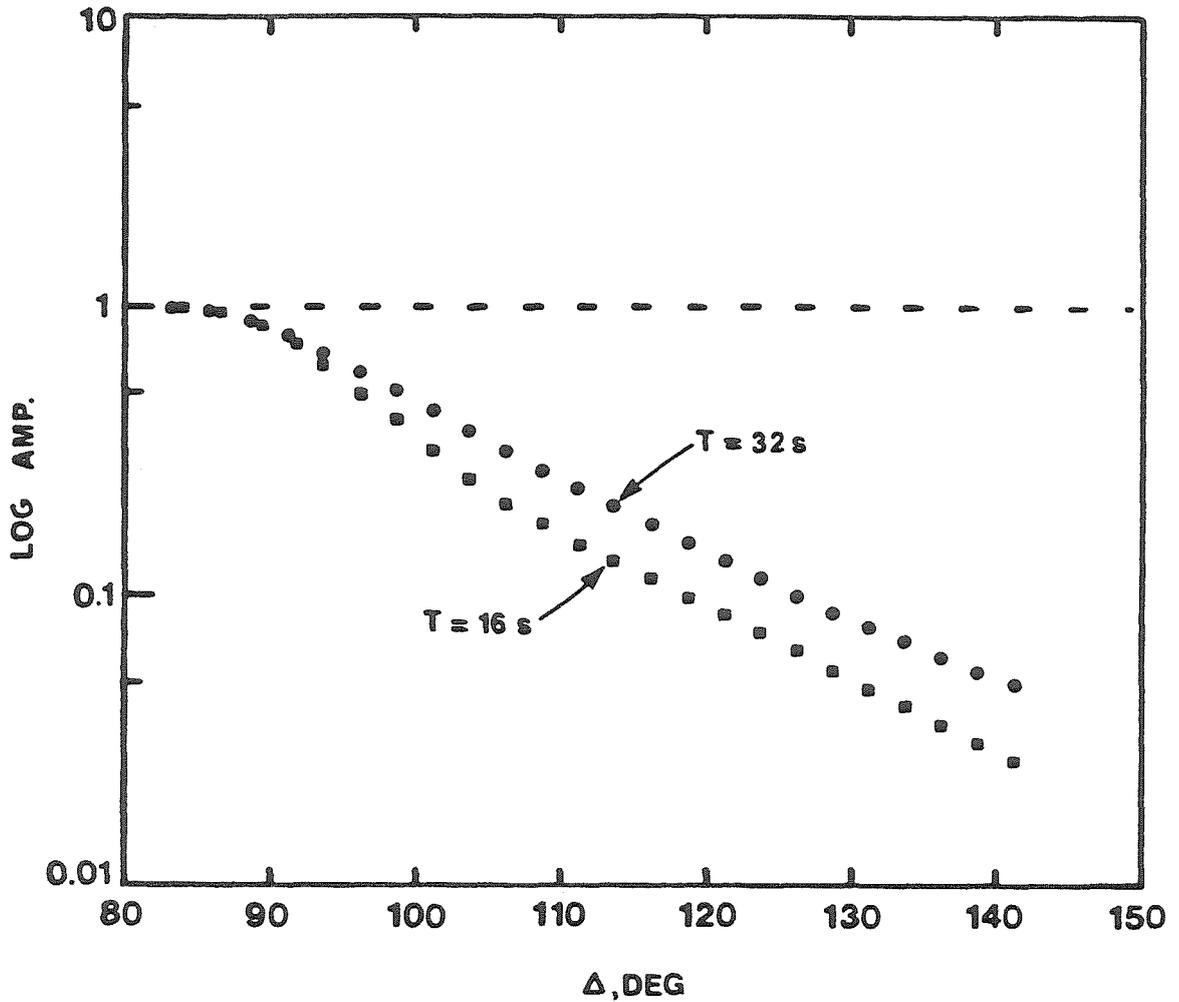
Figure 3.2. The decay of P wave amplitude into the shadow zone. The amplitude decay is frequency dependent, and curves for two representative periods are shown. The velocity model used is quite similar to model 1066B (see Part II).

In fact, the response is similar to an attenuation filter. Though calculating diffracted responses is non-trivial, synthetic seismogram methods have been developed which can generate the diffracted response for various structures at the core-mantle boundary. The structure in the lowermost ~200 km of the mantle is uncertain and may vary laterally. The currently allowed structural variations can affect the short period P waves in the shadow (see Part II), but fortunately long period P waves lack the resolution to be sensitive to these variations (see, e.g. Doornbos and Mondt,1979; Mula, 1981). The long period diffracted responses are invariably smooth and do not have later arrivals such as the upper mantle response. Examples of the amplitude decay into the shadow for two periods are shown in Figure 3.2. Observations of the long period amplitudes are quite scattered. Although we can make reasonable amplitude corrections based on theoretical results, the intrinsic scatter in observed amplitudes would always cause a factor of ~2 scatter in the earthquake moments. We have not included the diffracted response in the modelling of the data as the differences in the time functions between the events are so large as not to demand a formal correction of the diffraction effect.

3.3. Data

The main conclusion of this paper is readily apparent in the seismograms. In addition to comparing the seismograms of the great earthquakes, we include a "typical magnitude 8" event. The WWSSN long

period instruments were changed from a 30-100 to 15-100 (seismometer period - galvometer period) configuration in mid 1965. Therefore, to directly compare the seismograms, the reference magnitude 8 event must have occurred between 1962 and 1965. The only event that qualifies is the 1964 Niigata, Japan event, with M_W=7.6. This event has a thrust mechanism (see Figure 3.3), but is not a subduction zone earthquake. Figure 3.4 shows some of the records for the Niigata event. Note that the dominant effect of diffraction is to decrease the amplitude, with the waveform changing slowly to a smoothed version of the pre-shadow waveshape. Figure 3.5 shows the seismograms for the Kurile Islands event. The amplitudes of the Kurile Islands records are larger than those of the Niigata event. Also, the Kurile Islands records are composed of many pulses, as compared to the Niigata records which we consider as a single pulse, or event. Clearly the Kurile Islands earthquake is composed of many events of at least the same size of the Niigata event. The records of the Rat Islands event (Figure 3.6) appear similar to those of the Kurile Islands event, though of a larger amplitude.

The Alaskan records in Figure 3.7 are impressive, and two features of importance are easily seen: (i) the amplitude is substantially larger than that of the other events, (ii) the dominant period of the waveform is noticeably longer than the dominant period of the other events. (Representative records of the four earthquakes can be directly compared in Figure 3.8). In particular, the first pulse of the Alaska records remains positive for approximately 30 sec. This straightforward
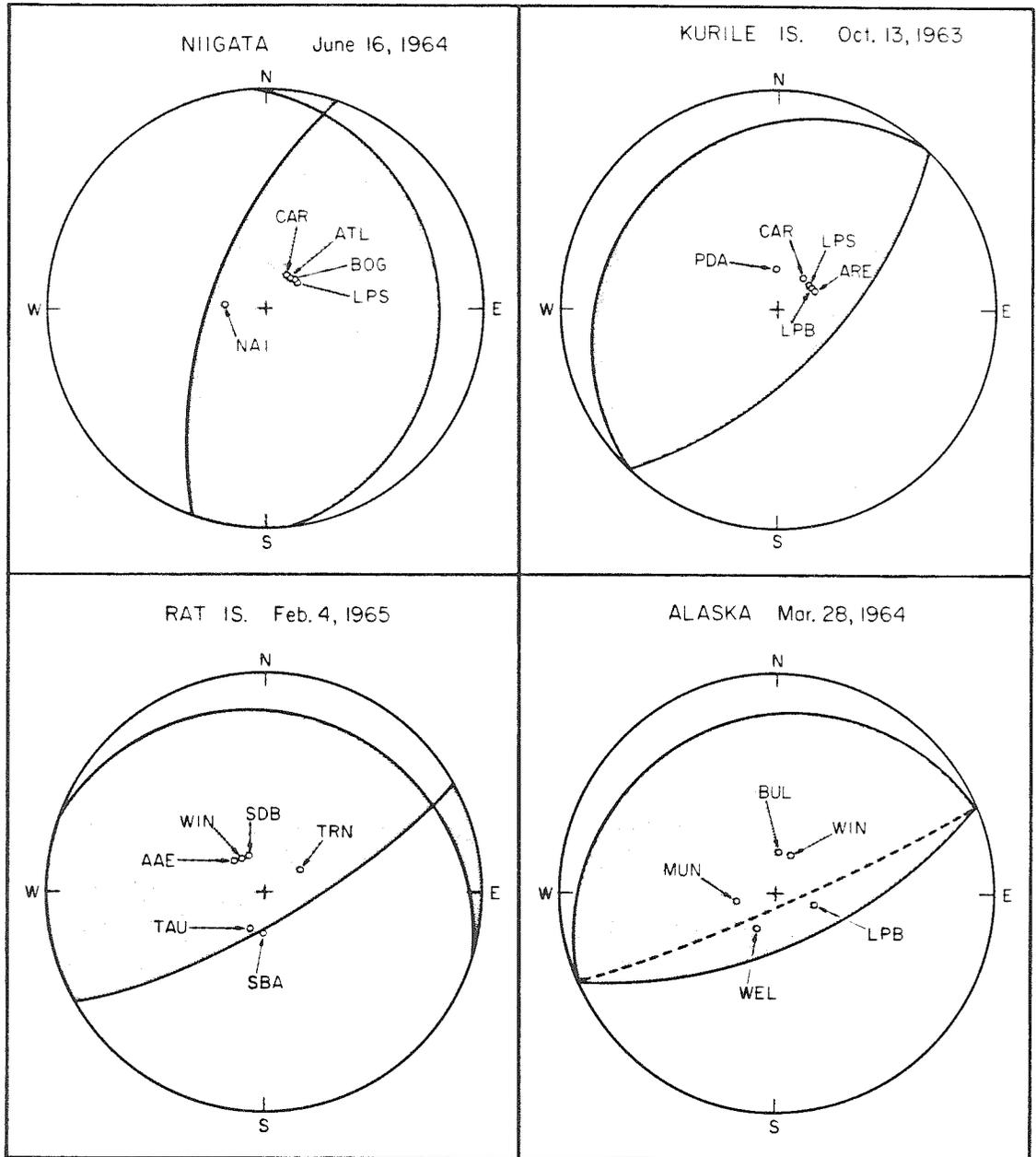
Figure 3.3. Focal mechanisms and locations of stations on the focal sphere. The shaded regions depict the compressional quadrants in the lower hemisphere projections. The mechanisms are from studies which used long period surface waves. The mechanism for Alaska is the "deep" (h = 70 km) solution of Kanamori (1970b), which is not consistent with first motions. The dotted plane is consistent with first motions.
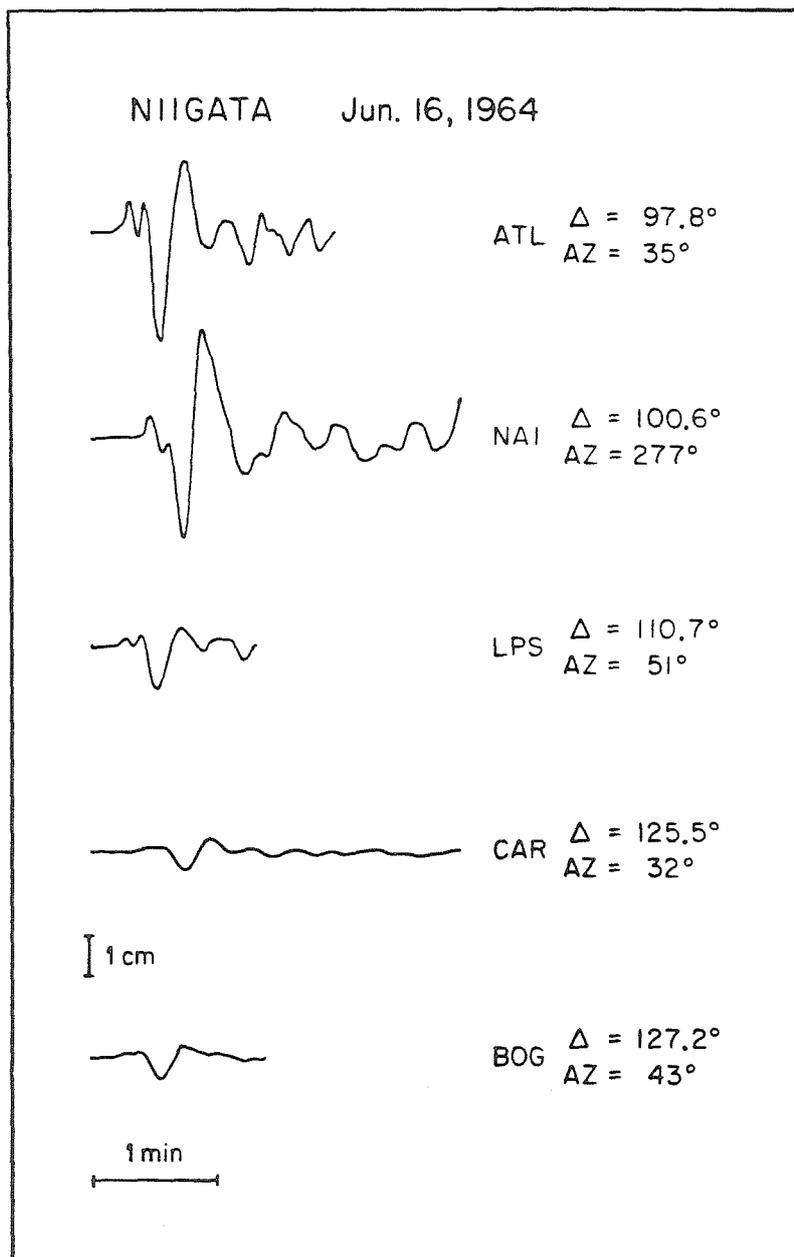
Figure 3.4. Representative seismograms of the Niigata earthquake. Notice the smooth amplitude decay into the shadow zone. This earthquake is characterized by a single pulse.
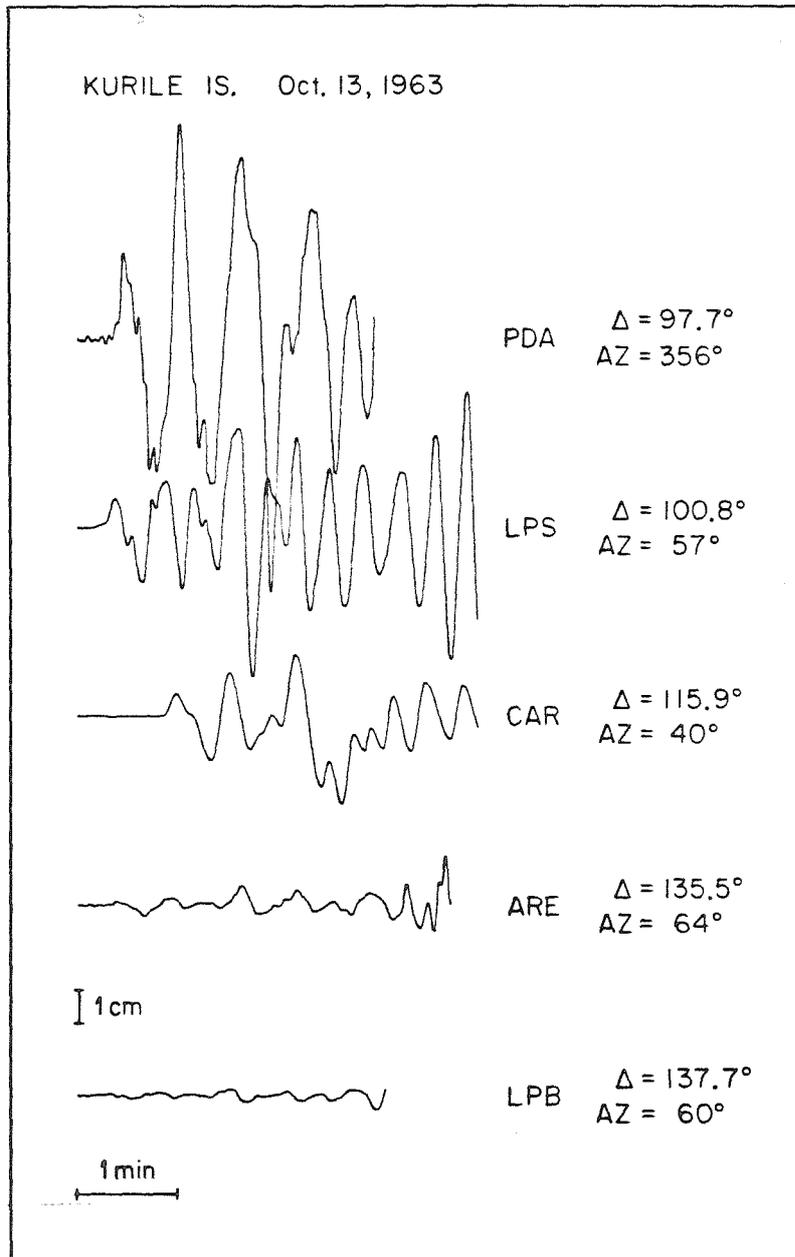
Figure 3.5. Seismograms of the Kurile Islands earthquake. There is a sequence of pulses and the overall amplitude is larger than the amplitude of the Niigata seismograms.
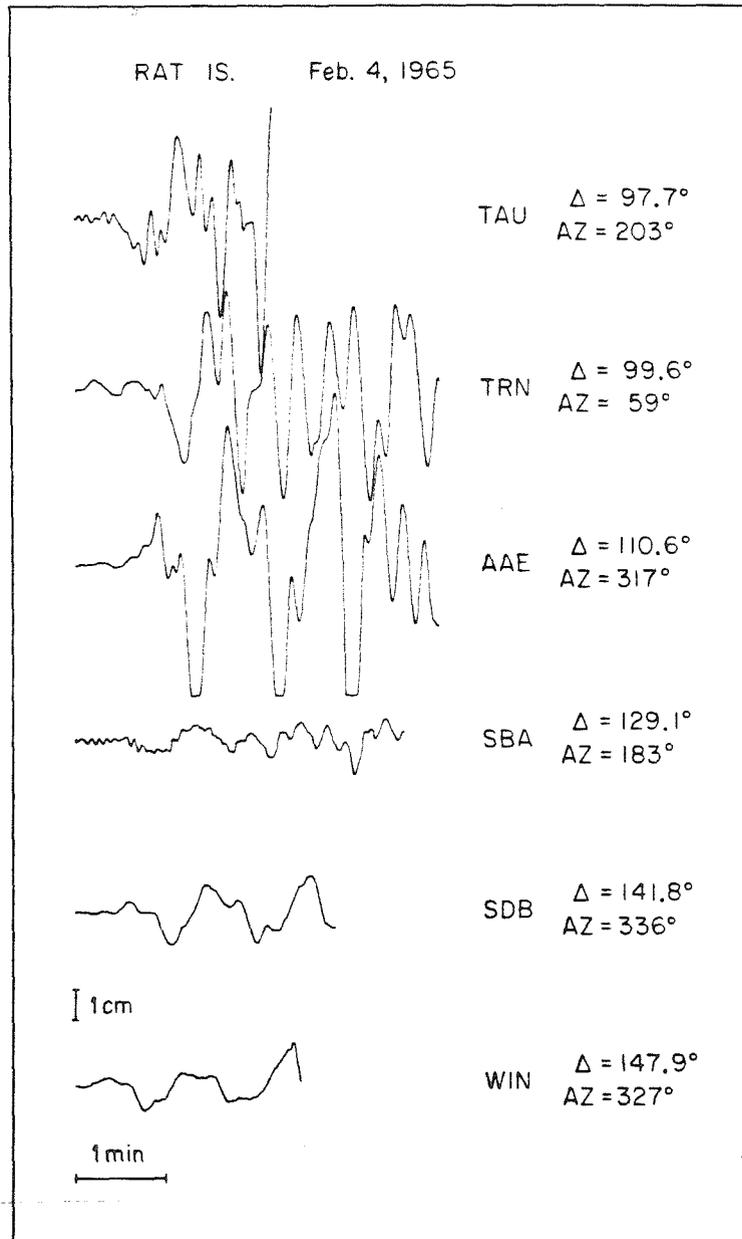
Figure 3.6. Seismograms of the Rat Islands earthquake. The amplitude is larger than the Kurile Islands amplitude (note the diminished amplitude scale), and the pulses appear to have a longer dominant period.

ALASKA   MARCH 28, 1964

WEL $\Delta = 106.5°$
       $AZ = 209°$

MUN $\Delta = 120.4°$
       $AZ = 259°$

BUL $\Delta = 139.0°$
       $AZ = 6°$

WIN $\Delta = 140.1°$
       $AZ = 22°$

I 1 cm

1 min
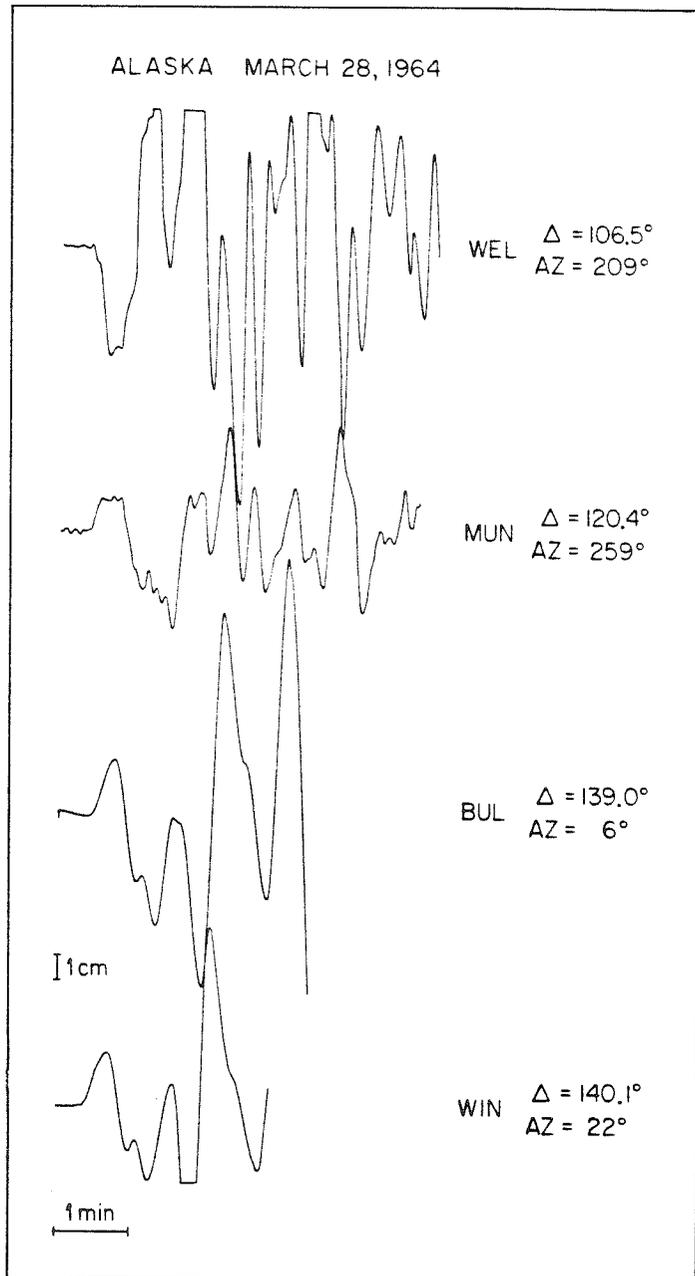
Figure 3.7. Seismograms of the Alaskan earthquake. The amplitudes are
very large and the dominant period is noticeably larger than for the
other earthquakes.
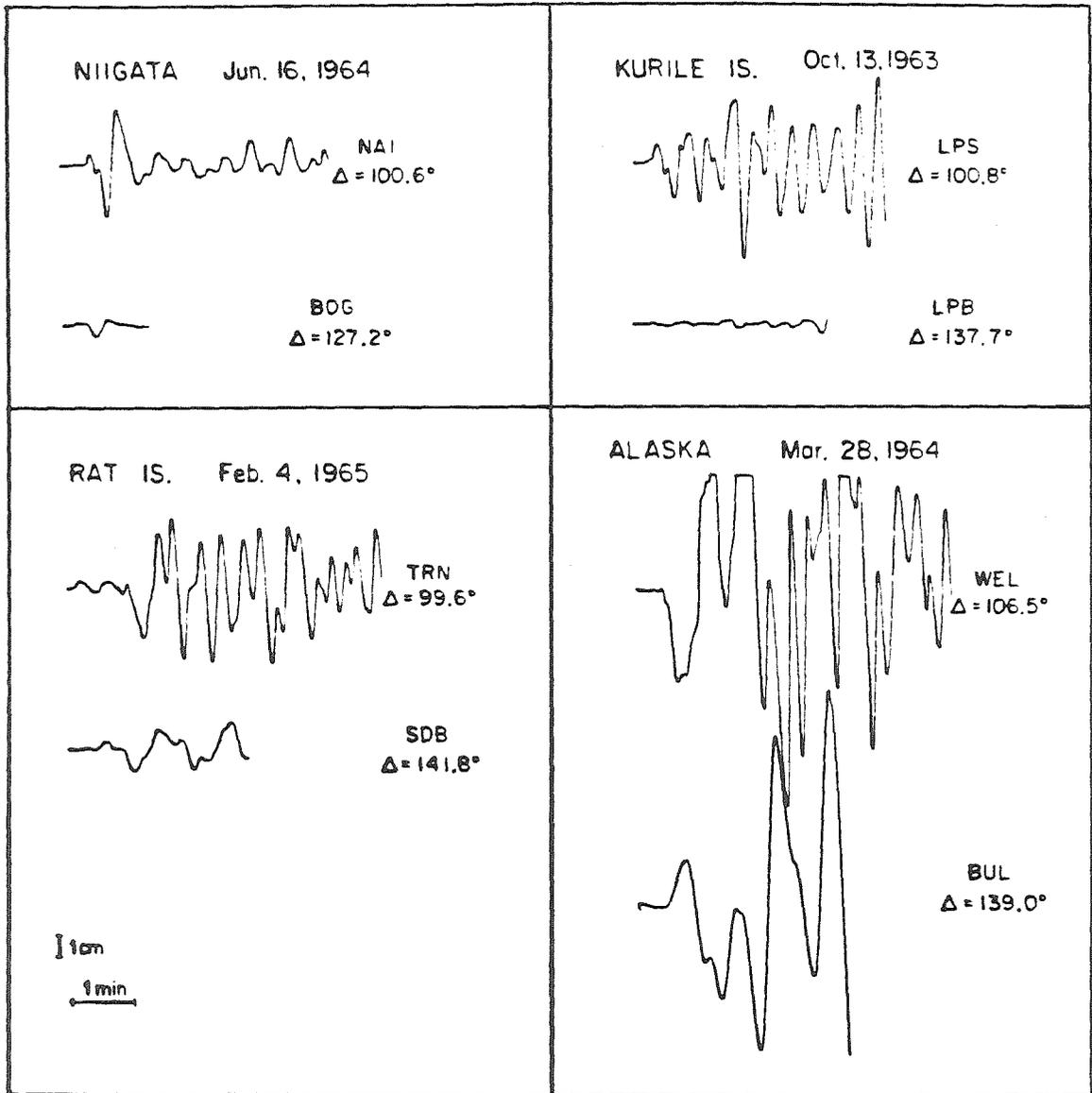
Figure 3.8. Two representative seismograms from each earthquake plotted at the same amplitude and time scales. The large size and long period nature of the Alaska seismograms is clearly evident.

observation has immediate and strong implications for the nature of the source time function and rupture process. Regardless of the source mechanism, it is quite difficult to force the 30-100 long period response to remain at one polarity for 30 sec. A ramp time function with a duration greater than 30 sec is not adequate. This is illustrated in Figure 3.9 where the BUL seismogram is modelled using two different time functions, a ramp and a half-cosine. Note that the fault orientation and source depth used in (3.9a) and (3.9b) are quite different. These two focal mechanisms are both acceptable in terms of the long period surface waves (Kanamori, 1970b), though the deeper mechanism is not consistent with the first motions (see Figure 3.3). Despite the uncertainty in the focal mechanism, the broad first pulse of the Alaskan records apparently requires a time function that curves upward for at least 30 sec. Also, the time function must continue as a smooth function, as any abrupt jump would cause a very large, sharp pulse that is not observed (for the first 3 minutes of the rupture). It is important to notice that this long period character exhibited by BUL is not just a consequence of diffraction. All of the Alaskan records show the same initial pulse width, including the LPB(NS) record at Δ=97.8 deg, where the diffraction effect is minimal (see Figure 3.10). The time function required at LPB is similar to that at BUL. Based on the time functions in Figure 3.10, it seems that the moment released in the first 90 seconds of the Alaskan earthquake is greater than the total moment of the Kurile Islands earthquake.

The main conclusion is already apparent. Based on a direct
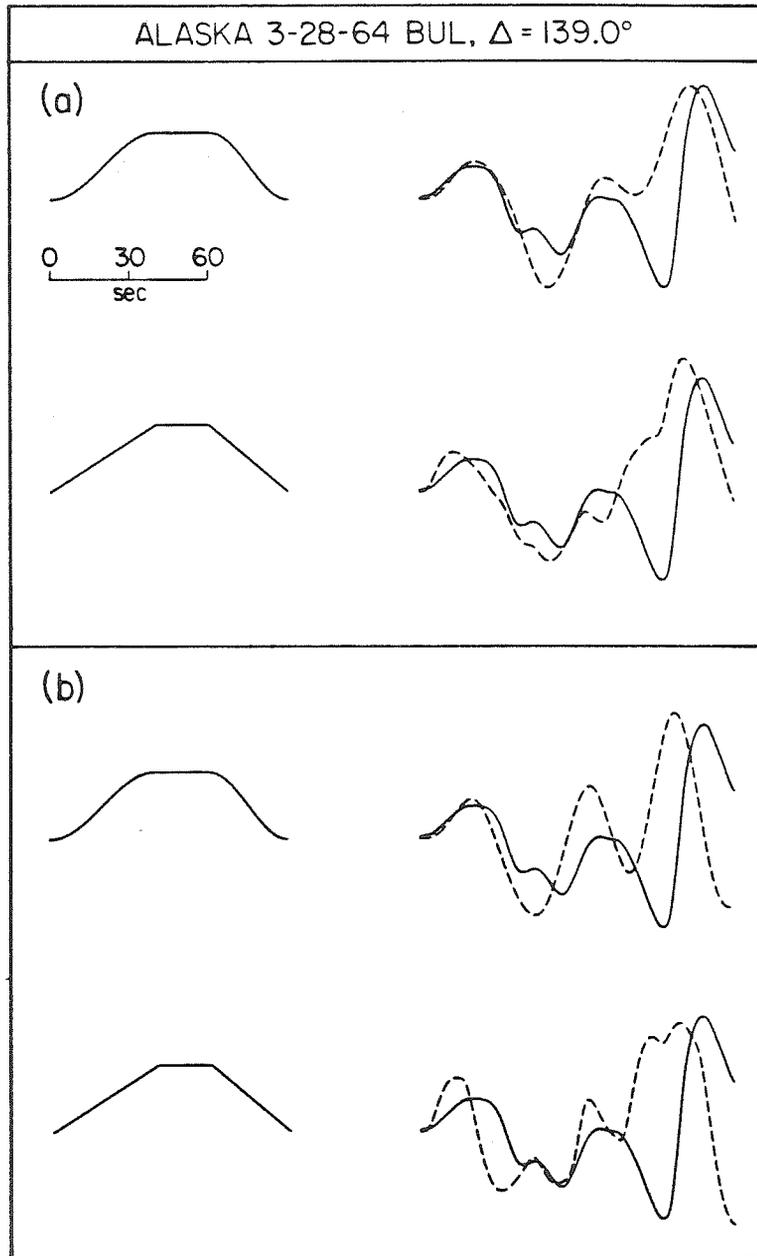
Figure 3.9. Forward modeling of the Alaskan record at BUL. The time functions are on the left, the observed seismogram is the solid trace, and the synthetic seismograms are the dashed traces. The width of the first peak of the seismogram cannot be matched by a time function with a simple ramp. Some type of "rising-ramp", such as a half-cosine, is required by the first part of the seismogram regardless of the mechanism used (3.9(a) and 3.9(b) show the synthetics for the deep and shallow mechanisms of Kanamori, 1970b).
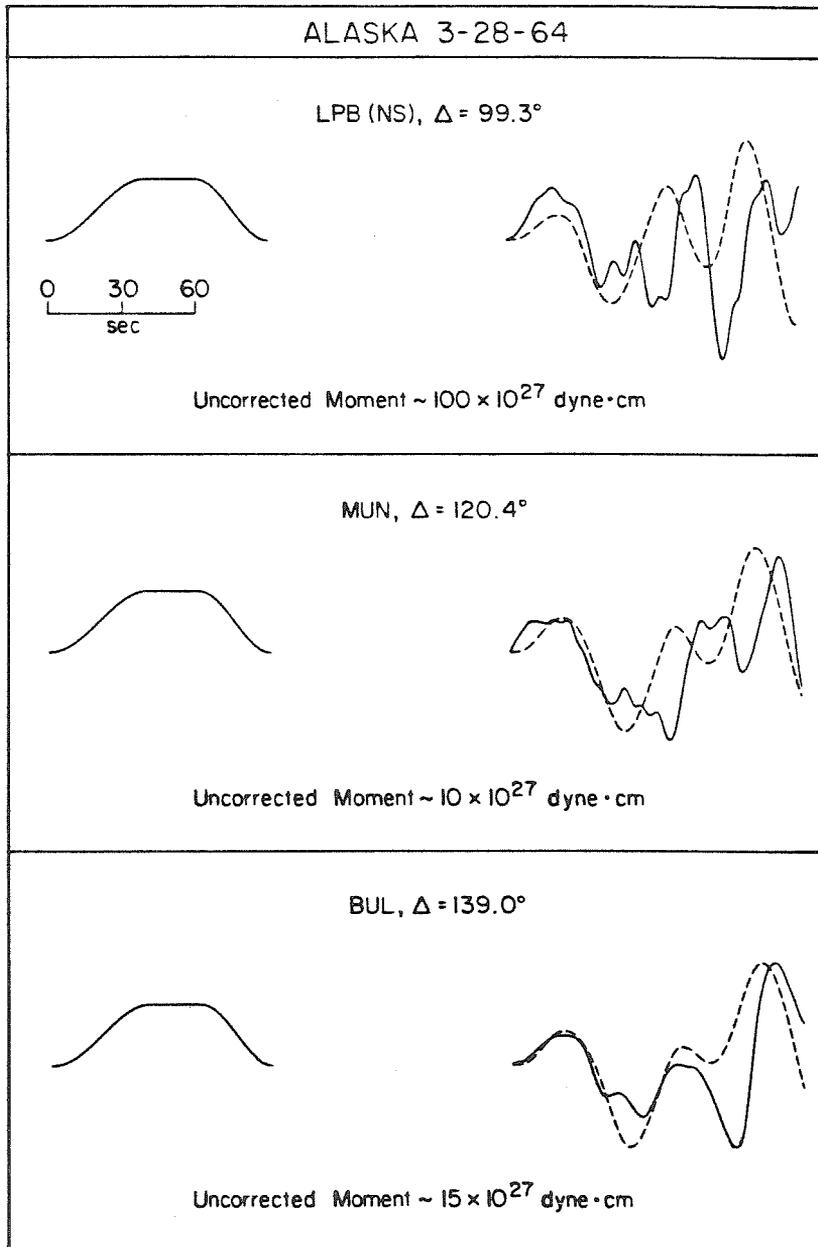
Figure 3.10. The Alaskan time function determined by forward modeling. The half—cosine time function used in Fig. 3.9(a) is applied to the seismograms recorded at MUN and LPB (NS component). The width of the first peak is not an effect of diffraction, hence a rising—ramp time function is required at all stations.

comparison of the seismograms combined with some simple forward modelling, the initial rupture of the Alaskan earthquake was a slow, smooth rupture with an effective rise time of greater than 60 sec, as compared to the Niigata event with a total rupture time of ~20 sec. As will be discussed later, this very long rise time translates into a source area of 140 to 200 km for the Alaskan asperity that initiated the earthquake.

To quantify the above conclusion, we will need to determine the source time functions for each of the events. A method for determining source time functions is presented in the next two sections.

3.4 Body Wave Seismograms of Earthquake Sources

Synthetic seismograms for body waves at teleseismic distances can be constructed as $s(t)=I(t)*g(t)*m(t)$, where $*$ denotes convolution, $s(t)$ is the seismogram, $I(t)$ is the instrument impulse response, $g(t)$ contains the source geometry and the earth's impulse response, and $m(t)$ is the source time function, i.e. the faulting behavior as a function of time. Assuming the average fault geometry is known, $m(t)$ is then obtained by inverting the convolution operation. This can be done in either the frequency domain or directly in the time domain. We will use a time domain method.

Point source representation

The earthquake source is usually considered as a shear dislocation across the fault surface. If the region of displacement discontinuity is small (i.e. source dimension is smaller than the wavelength of interest), then the earthquake source can be considered as a point source. The source displacement can be characterized by a system of a force couples, conveniently grouped into the moment tensor M_{ij}, $i,j=1,3$ (see, for example Aki and Richards, 1980). If we restrict our attention to a shear dislocation across a planar surface, then the moment tensor can be reduced to a double couple of some orientation. For the teleseismic (station distance greater than source dimension) body wave problem, the double couple source is represented by the radiation pattern which gives the displacement amplitudes of P or S waves leaving the source (for details, see Langston and Helmberger, 1975; Kanamori and Stewart, 1976). For a double couple point source, the source time function is the rate of moment release and has the same time dependence at all azimuths. The effect of propagation through the earth is then given by the appropriate point force Green's function. Thus, the direct P wave recorded at distance Δ and azimuth is given by

$$S(t,\Delta,\phi)=I(t)*G_p(t;\Delta,h)*R_p(\Delta,\phi,h)\dot{M}(t),$$

(3.1)

where $I(t)$ is the instrument response (includes the free surface effect), $g_P(t;\Delta,h)$ is the Green's function, $R_P(\Delta,\phi)$ is the radiation amplitude, and $\dot{M}(t)$ is the moment rate. When the station distance is in the range where geometric ray theory is applicable, 30 to 90 deg, the Green's function is simply the attenuation operator multiplied by the geometric spreading factor and source constants, and delayed by the travel time: $G_P(t;\Delta,\phi)=g(\Delta,h)Q(t)*\delta(t-T_P)$. For shallow earthquakes, the reflections from the surface, pP and sP, arrive quite soon after the direct P wave and need to be incorporated into the synthetic. The seismogram is then,

$$
\begin{aligned}
S(t,\Delta,\phi)=I(t)*[\ &R_P(\Delta,\phi,h)\delta(t-T_P)+r_{pP}R_{pP}(\Delta,\phi,h)\delta(t-T_{pP}) \\
&+r_{sP}R_{sP}(\Delta,\phi,h)\delta(t-T_{sP})\ ]*g(\Delta,h)Q_P(t)*\dot{M}(t) \quad ,
\end{aligned}
\tag{3.2}
$$

where R_{pP} gives the amplitude for the upgoing P wave, r_{pP} is the surface reflection coefficient, T_{pP} is the total pP travel time, and similarly for sP. Note that the above expression can be shortened to $s(t)=I(t)*g(t;\Delta,\phi,h)*m(t)$ where $g(t;\Delta,\phi,h)$ is referred to as the "half space" Green's function for a particular mechanism. With the above expression, it is quite easy to construct synthetic seismograms for a given focal mechanism and source time function ($m(t)$).

Finite source dimension

As we will model earthquakes with a fault length of 100 km or more, the point source approximation is not valid. For a finite source, we need to integrate the contributions from point sources distributed over the fault plane,

$$S(t,\Delta,\phi)=I(t)*\iint_A G(t;\Delta,\phi,h)*\mu\dot{D}(t)\ dA\ ,\tag{3.3}$$

where $dM=\mu D dA$, $\dot{D}$ the time derivative of the particle displacement, and μ is the shear modulus in the source region. $g(t;\Delta,\phi,h)$ will vary over the fault plane, and yields the directivity effect due to the variation in Δ,ϕ, and h across the fault plane. Additionally, $g(t;\Delta,\phi,h)$ could vary if the focal mechanism changed over the fault surface. Usually the variation in $g(t)$ over the fault surface is slight, so that an average $g(t)$ can be taken outside of the integral,

$$S(t,\Delta,\phi)=I(t)*\overline{G}(t;\Delta,\phi,h)*\mu\iint_A \dot{D}(t)\ dA\ ,\tag{3.4}$$

which has the same form as the point source expression except that the source time function is now $\mu \iint \dot{D}(t)dA$. To interpret this, it is convenient to use a rupture front sweeping over the fault area. This allows a change of variable from dA to $d\tau$, where $A(\tau)$ gives the location of the rupture front as parameterized by the rupture time, τ. Thus,

$$S(t,\Delta,\phi)=I(t)*\overline{G}(t;\Delta,\phi,h)*\mu \int_0^\infty \dot{D}(t,\tau)\frac{dA}{d\tau}\ d\tau \quad . \qquad (3.5)$$

If the particle displacement history is the same everywhere on the fault, simply delayed by the rupture time, then $\dot{D}(t,\tau) \rightarrow \dot{D}(t-\tau)$, and

$$S(t,\Delta,\phi)=I(t)*G(t;\Delta,\phi,h)*\mu \int_0^\infty \dot{D}(t-\tau)\dot{A}(\tau)\ d\tau \quad . \qquad (3.6)$$

The integral is now in the form of the convolution integral. The time function $m(t)$ is the convolution of the particle velocity and the rate of increase of the faulted area. Two simple examples are shown in Figure 3.11, unilateral and circular rupture. For large earthquakes, the duration of rupture across the fault plane is longer than the characteristic time of particle displacement, hence the duration of the time function indicates the rupture time which is the characteristic fault length divided by the rupture velocity. In some cases, the
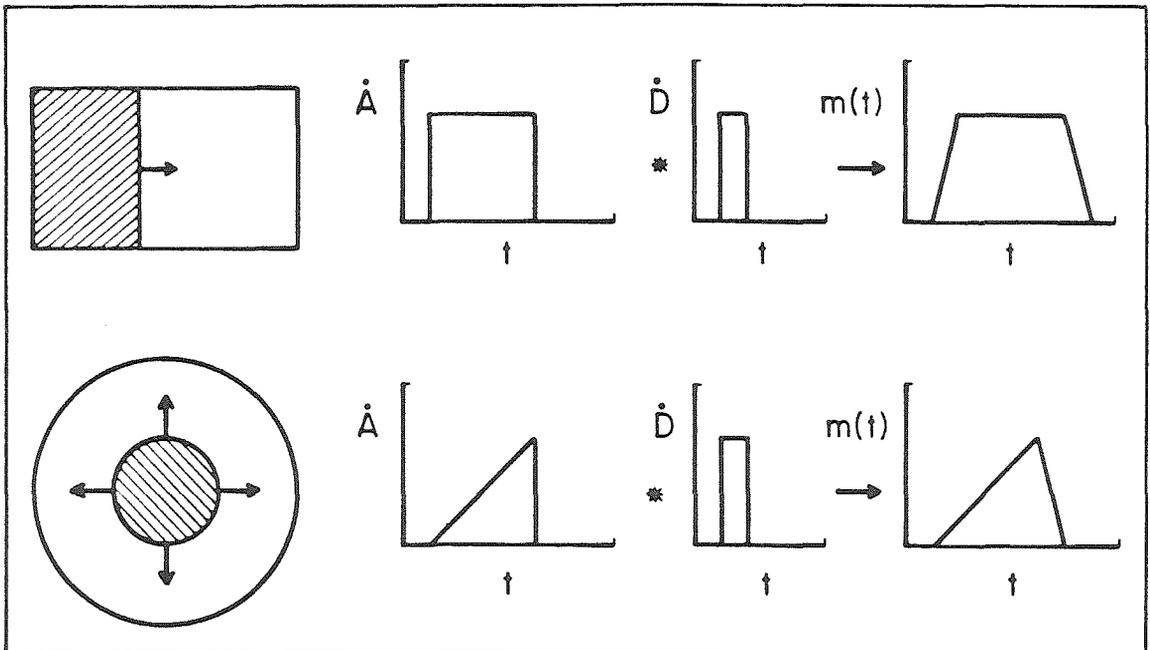
Figure 3.11. Two different rupture modes, unilateral and circular. The time function for unilateral rupture is depicted in the top row. At the far left, a rectangular fault is shown with the rupture front sweeping to the right. When the particle velocity is the same everywhere on the fault, the time function, m(t), results from the convolution of area rate with particle velocity. As area rate is characterized by a boxcar function, the time function will have the shape of a trapezoid. For the case of circular rupture (shown on bottom row), a ramp function characterizes the area rate, leading to a sawtooth time function for an abrupt termination of rupture.

observed shape of the time function indicates the mode of rupture, e.g. circular or unilateral.

Returning to equation (3.6), the directivity effect can be included by allowing $dA/d\tau$ to vary with azimuth to station. Directivity is usually ignored for P waves as the total effect is only a few seconds for fault lengths less than 100 km.

To determine the source time history by using equation (3.6), it is necessary to choose an average Green's function. The simplest choice is to use the half-space Green's function, selecting some average source depth. As the half-space Green's functions vary most rapidly with depth, perhaps a finite source distributed over a depth interval is not well represented by a fixed depth Green's function. Figure 3.12 compares a fixed depth Green's function and Green's functions for distributed sources. The finite Green's function appears to be a filtered version of the point source Green's function. Hence, the synthetic seismograms for the point source and distributed source would be quite similar after convolution with a long period time function. We have used a fixed depth Green's function in the following analysis, and we will later show a comparison of time functions obtained from fixed depth and distributed source Green's functions. We do not imply that the Great Alaskan earthquake is a point source, as the source finiteness will be exhibited in the time function through the $dA/d\tau$ contribution in the integral. Choosing an average Green's function simply means that we are not attempting to model the details of the time function, that is we are mainly interested in the gross features of the rupture histories.
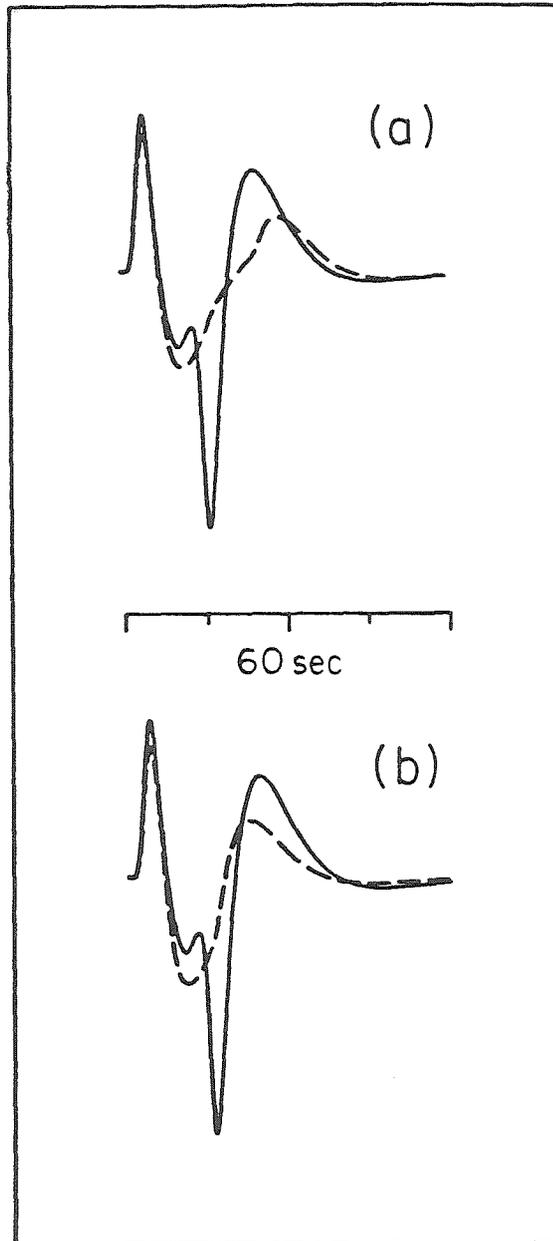
Figure 3.12. Half-space Green's functions for a point source and distributed sources. The solid trace in (a) and (b) is the point source Green's function at BUL for the Alaskan earthquake with h = 30 km. The dashed trace in (a) is the Green's function for a source distributed from the surface to a depth of 50 km, while the dashed trace in (b) is for a source between the surface and 30 km. The WWSSN 30-100 instrument is used. The distributed source Green's functions appear as filtered versions of the point source Green's function.

3.5. Time Function Deconvolution

In the following section, the methodology of deconvolution is discussed. If one is not concerned with critically examining the seismic moments or the construction of the time functions, then this section can be skipped. The time functions can then be regarded as lucky guesses, and evaluate their worth by whether the associated synthetic seismograms fit the data and if the time functions determined for individual stations are similar.

The seismogram is modelled as $s(t)=g(t)*m(t)$, where $g(t)$ now includes the instrument, propagation effects, and the assumed source mechanism, and $m(t)$ is the effective far-field time function. Determining a satisfactory $m(t)$ by trial-and-error modelling for a large number of seismograms is not practical as we anticipate rather complicated time functions. It is preferable to deconvolve $m(t)$ directly from the observed seismogram. Deconvolution is a subject of general interest in signal analysis and is routinely used in seismic reflection data processing. As convolution is a linear operation, it is surprising that no generally useful computational technique is available. The primary reason for this is that deconvolution is unstable, with respect to the reliabilty of $m(t)$, for most applications and this problem is handled in different ways. We will review the known properties of deconvolution, then discuss our particular application including the reliability and resolution of the solutions.

Frequency domain deconvolution

Though $m(t)$ will be deconvolved in the time domain, the basics of frequency domain deconvolution are presented due to the conceptual simplicity of this approach. Note that the effects of discrete sampling will not be treated (which occur when using the Fast Fourier Transform algorithm).

Frequency domain deconvolution is simple because convolution in the time domain is equivalent to multiplication in the frequency domain. Let $S(f)$, $G(f)$, and $M(f)$ be the Fourier transforms of $s(t)$, $g(t)$, and $m(t)$ respectively. Then, the Fourier transform of the equation $s(t)=g(t)*m(t)$ is $S(f)=G(f)M(f)$ for "reasonable" functions $g(t)$ and $m(t)$ (requires changing the order of integration). Convolutions are commonly performed by complex multiplication in the frequency domain, e.g. $s(t)=F^{-1}(G(f)M(f))$, where F^{-1} denotes the inverse Fourier transform. To retrieve $M(f)$, it seems that we can simply divide by $G(f)$, $M(f)=S(f)/G(f)$, and $m(t)$ is then the inverse Fourier transform of $M(f)$, $m(t)=F^{-1}(S(f)/G(f))$. Unfortunately, problems arise at the step of dividing by $G(f)$. If $G(f)=0$ at particular values of f, then $1/G(f)$ is unbounded and $M(f)$ can be unbounded or finite depending upon the details of $S(f)$ at these points. Thus, $M(f)$ is unstable with respect to infinitesimal variations in $S(f)$ at the zeroes of $G(f)$. Practically, this instability of $M(f)$ appears whenever $G(f)$ and $S(f)$ are smaller than the scatter in $S(f)$.

The difficulties with 1/G(f) are resolved in two ways: (i) a filter is used to smooth 1/G(f), which results in a time limited m(t), (ii) M(f) is simply set to some predetermined value where G(f)≡0, which results in a band limited m(t) and possibly spurious frequency components in m(t) that are not resolved. Note that for g(t) a purely real function, $F^{-1}(1/G(f))$ will also be purely real. If the frequency domain filter is real and even, then $m(t)=F^{-1}(S(f)/G(f))$ will be real though probably non-causal.

Time domain deconvolution

A more general approach is that of the Backus-Gilbert geophysical inverse theory (Backus and Gilbert, 1970). Given the equation,

$$s(t)= \int_{0}^{\infty} g(t-t') \, m(t') \, dt' \quad , \qquad (3.7)$$

relating the model, m(t), to the data, s(t), we ask: what properties of m(t) can be determined from observations of s(t)? In the following discussion it is assumed that g(t) is known. The uncertainties in m(t) arise from the properties of g(t) and inadequacies and errors in s(t). If a solution exists, it will be unique if the only solution to the homogeneous equation, $\int g(t-t')m_N(t')dt'=0$, is the trivial solution. In frequency domain, this equation is transformed to $G(f)M_N(f)=0$. This

relation is satisfied by a non-trivial $M_N(f)$ if $G(f)$ has zeroes. Hence, $M_N(f)=0$ except at values of f where $G(f)=0$, then $M_N(f)$ is arbitrary. This is the same difficulty as discussed above. Let $m_1(t')$ be a solution to equation (3.7), then contributions from the null space of operator $g(t-t')$ can always be added, that is $m_2(t')=m_1(t') + m_N(t')$ and $m_2(t')$ satisfies equation (3.7) also. If one guessed a trial-and-error solution $m(t')$ that fit the data, there could be arbitrary unresolved contributions from the null space of $g(t-t')$. A stable deconvolution method must control the null space contributions to the solutions.

To construct solutions to equation (3.7), it is necessary to discretize $s(t), g(t)$, and $m(t)$, thereby changing the integral into the following summation (for the simplest case of equal sampling),

$$s_i = \sum_{j=1}^{i} g_{i-j+1} \, m_j \; . \tag{3.8}$$

However, the Backus-Gilbert approach does not immediately discretize the entire problem. Instead, this method assumes that the data are available at only discrete values of t, but $m(t')$ and $g(t-t')$ are still considered as functions of t', hence

$$s(t_i) = \int_0^\infty g_i(t_i - t') \, m(t') \, dt' \quad . \tag{3.9}$$

Discretizing s(t) guarantees an infinity of solutions for m(t). The Backus–Gilbert method states that although m(t) cannot be uniquely determined at a point, say t_0, a localized average of m(t) around t_0 can be determined. That is, $m(t_0)$ is the average m in the interval $t_0 - y/2$ to $t_0 + y/2$ where y is the averaging length. The trade-off between resolution and stabilty can be stated as follows: the average m, $m(t_0, y)$, is less reliable for a smaller y (higher resolution). The localized average can be written as

$$\bar{m}(t_0) = \int_0^\infty R(t_0, t') \, m(t') \, dt' \quad , \tag{3.10}$$

where $R(t_0, t')$ is a unimodular averaging kernel. We wish to construct $\bar{m}(t_0)$ from a linear combination of the data,

$$\bar{m}(t_0) = \sum_{i=1}^{N} a_i s_i \quad . \tag{3.11}$$

Combining equations (3.9), (3.10), and (3.11), R is given by

$$R(t_o, t') = \sum_{i=1}^{N} a_i\, g_i(t_i - t') \quad .$$

(3.12)

As $R(t_o, t')$ tends toward a delta function, $\bar{m}(t_o)$ tends toward $m(t_o)$, i.e. perfectly resolved. The stability of $m(t)$ is derived from equation (3.11) as $\mathrm{var}(m(t_o)) = \sum a_i^2(t_o)\mathrm{var}(s_i)$. Thus, one needs to select a set of a_i so that R is delta-like yet the sum of a_i^2 is small. For a given level of resolution (delta-ness), the set of a_i that minimize the variance of $m(t)$ can be found (see Backus and Gilbert, 1968). A subjective choice must be made regarding the acceptable levels of resolution and variance.

There will be a minimum averaging length y_{min} associated with maximum resolution. Thus, the sampling of $m(t)$ should not be less than y_{min}. The sampling interval that we use for $m(t)$ is larger than y_{min}. Typical values of the sampling are: 1 sec sampling of the seismogram and 4 to 6 sec sampling of the time function.

The $g(t, t')$ function includes the instrument, which causes the spectrum to tend toward zero at very long periods as well as at short periods. While the uncertainties at short periods can be avoided by the sampling of $m(t)$, the lack of a zero frequency component is bothersome. Because of the physical interpretation of the time function, it should take on positive values (unless the focal mechanism is drastically

incorrect). Thus, after deconvolving the time function, a half-sine wave component with a period twice the time function duration is added so as to make the time function predominately positive. The duration of $m(t)$ is typically 180 sec, so the addition of the half-sine component flattens the spectrum from T=360 sec to zero frequency.

Generalized inverse

We have deconvolved $m(t)$ by solving equation (3.8) using the damped generalized (Lanczos) inverse (see Lanczos, 1961; Jackson, 1972; Wiggins, 1972). With the seismogram sampled more frequently than the time function, equation (3.8) is an overdetermined system, $A_{ij}m_j = s_i$. Smoothing conditions can be explicitly included by adding equations that set the first differences of m_j to zero,

$$
\gamma
\left\{
\begin{array}{c}
A \\
\text{-- -- -- -- --} \\
\begin{matrix}
1 & -1 & & & & \diagdown & \diagup \\
 & 1 & -1 & & & & 0 \diagdown \\
\diagdown & & 1 & -1 & & \diagdown \\
0 & & & \ddots & \ddots \\
\diagup \diagdown & & & & \ddots & \ddots
\end{matrix}
\end{array}
\right\}
\underline{m} = \underline{S} \quad .
\tag{3.13}
$$

The parameter γ weights the smoothing conditions relative to fitting the data. In practice, it is not necessary to enforce smoothness as the sampling of $m(t)$ results in adequately smooth solutions.

There are basically two methods available for constructing a stable inverse to A, The spectral (eigenvalue) method and the damped Lanczos

inverse. We have used the Lanczos inverse as it is remarkably simple and computationally efficient. The undamped Lanczos inverse is,

$$A^* = (A^T A)^{-1} A^T \quad , \tag{3.14}$$

where superscript T denotes the transpose of A. This formula can be derived by minimizing the data mismatch in a least squares sense, i.e. $\min(e^T e)$ with $e_i = A_{ij} m_j - s_i$. With the above formula, only the $A^T A$ matrix needs to be inverted. This square matrix is symmetric and positive definite, hence the inverse exists for rank equal to j. As the Lanczos inverse results in the least squares estimate for m_j, it provides an exact answer if any exist. Also, it gives the maximum resolution of $m(t)$. Let some inverse of A be A^*. Then our estimate of the model (m_j) is,

$$\bar{m}_j = A^*_{ji} A_{ik} m_k = A^*_{ji} s_i \quad . \tag{3.15}$$

The resolution matrix is $R = A^* A$, and is similar to the averaging kernel in the continuous case. The resolution matrix for the Lanczos inverse is the identity matrix. The Lanczos inverse can produce an unstable estimate for $m(t)$ if the $A^T A$ matrix is "almost" singular. This

situation is analogous to allowing null space contributions. Poorly resolved components of the solution can be eliminated by partially minimizing the length of the solution vector, m_j. That is, instead of minimizing the data mismatch, minimize the quantity $[(1-r)e^T We + rm^T Vm]$ where $W^{1/2}$ weights the data (usually, the diagonal elements of W are $1/var(s_i)$), V represents an a priori estimate of the model variance, and r is an adjustable parameter. This parameter indicates the relative weight between fitting the data and forcing a "shorter" solution. The damped Lanczos inverse is,

$$A^* = (A^T A + dV)^{-1} A^T \quad , \tag{3.16}$$

where d is the damping parameter, $d=r/(1-r)$. Experimenting with different assumptions for V, it was found that various weighting schemes produce solutions generally similar to the solution when V is the identity matrix. Note that for non-zero d, the resolution matrix will no longer exactly equal the identity matrix.

As d is an important parameter that strongly affects the solution, we examine the change in the resolution and inverse matrices associated with a change in d. The Alaskan earthquake record at BUL is used. Figure 3.13 shows the solution, corresponding synthetic seismogram, observed seismogram, and the R and A^* matrices for a very small value of d, $d=10^{-5}$. The slight damping allows long period components off the
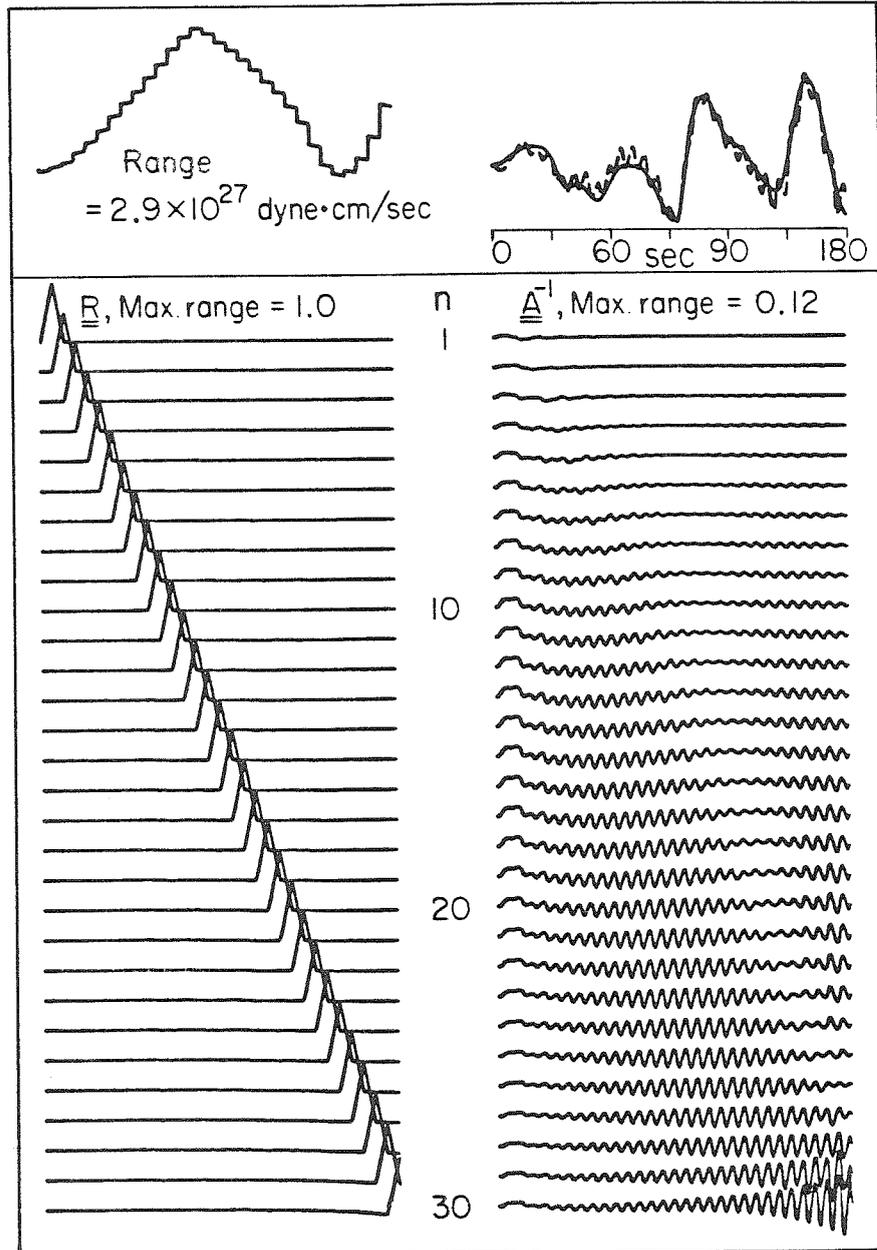
Figure 3.13. The deconvolved Alaskan time function for the damping constant = 10^{-5}. The solid trace at upper right is the seismogram recorded at BUL. The deconvolved time function is at upper left and the corresponding synthetic seismogram is the dashed trace. The time function is composed of 30 points with a spacing of 6 sec. The rows of the resolution matrix are shown on the left, and the rows of the inverse matrix are shown on the right.
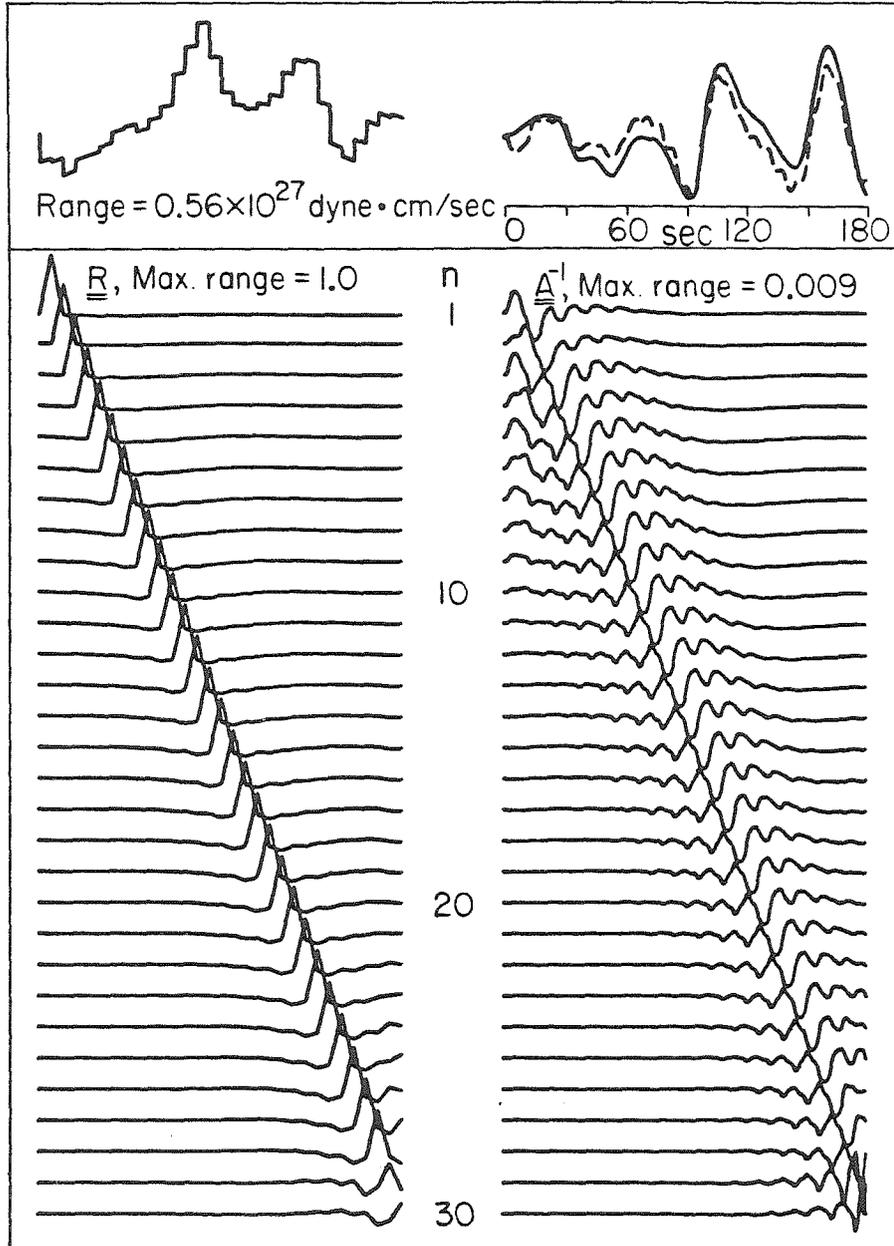
Figure 3.14. The deconvolved Alaskan time function for the damping constant = 10^{-2}. Same as Figure 3.13 except that the damping constant used in constructing the inverse is three orders of magnitude larger. The resolution is degraded only slightly, but the inverse matrix is severely attenuated.

edge of the instrument pass-band to appear in the solution. The resolution matrix is essentially the identity matrix. The rows of A^* show how the seismogram is weighted to obtain the model values, and correspond to the sets of a_i discussed previously in the Backus-Gilbert view. Note that a particular value of m_j depends upon almost the entire seismogram. Recall that the variance of m_j is related to the data variance by the squared length of the row. Hence, the variance of the time function increases with time for a uniform data variance. For example, if the data variance is 25% of the maximum amplitude, the variance of the time function in Figure 3.13 would be .03% of the maximum amplitude for the first time point and 8% for the last time point. Figure 3.14 shows the effects of increasing d to 10^{-2} (the largest d value used in the analysis). The resolution is not significantly affected, indicating that the sampling interval is large such that the elements of the time function will always be formally well resolved. Increasing the damping has had a more significant effect on the inverse matrix, and consequently on m_j. The peak to peak amplitudes of the rows A^* are reduced by an order of magnitude. Also, the high frequency component seen in Figure 3.13 has been seriously attenuated. Notice that the solution is now composed of a more local weighting of the seismogram. The relative variance of the model is smaller with the larger damping, for 25% data variance the model variance is .1% at the first point and .05% at the last point.

These two figures also indicate the model features that are demanded by the data. Certainly the summed moment is highly sensitive

to the value of d. With $d=10^{-2}$, the amplitude of A^* is reduced by an order of magnitude and the total moment of the time function is an order of magnitude lower. The addition of the half-sine component does not restore the moment to the value for $d=10^{-5}$. However, notice that the character of the first part of the time function is retained, in particular the smooth ramp with a 70 sec duration.

The procedure used to model the seismograms is to first deconvolve the time function with a large damping, $d=.01$. If the synthetic seismogram for this time function does not match the data satisfactorily, then the damping is reduced. Hence, we are determining minimum seismic moments.

Inadequacies in $g(t,t')$

One aspect of geophysical inverse theory that has not been well explored is the effect of errors in $g(t,t')$. In the previous development it has been assumed that $g(t,t')$ is the true $g(t,t')$. Aside from approximations in the theory used to derive $g(t,t')$, there may be errors in the assumed average focal mechanism and source depth.

If the errors in $g(t,t')$ are: random, uncorrelated, normally distributed, and each row of the A matrix ($g_i(t')$) has the same variance, then we can accommodate the variance of g_i in the same manner as the variance of s_i. Each equation would be weighted by $1/var(g_i)$ instead of $1/var(s_i)$. This procedure is intuitively acceptable as the unexplained randomness in s_i can be considered as random inadequacies in

g_i. However, the errors in $g(t,t')$ are probably not random. They could be highly correlated as errors in the focal mechanism can cause systematic effects. It is difficult to correct for systematic errors as the character of the systematic error must be known, but then it would not be an error.

The approach followed here is to simply demonstrate the size of possible errors, with particular attention given to the longer periods. For a clearer presentation the effects will be shown in the frequency domain. Let the observed seismogram be $S(f)=G_o(f)M_o(f)$ where $G_o(f)$ and $M_o(f)$ are the "true" Green's function and time function. We choose a Green's function, $G(f)$, and deconvolve $M(f)$ so that $M(f)=S(f)/G(f) = M_o(f)G_o(f)/G(f)$. If $G(f)$ is the true Green's function, then the true time function is recovered (here ignoring the deficiencies of $1/g$). However, if $g(t,t')$ does not equal $g_o(t,t')$, then M_o is contaminated by the factor (G_o/G). In Figure 3.15, this bias in spectral amplitudes is plotted for perturbations to the source description about a "standard" mechanism. The parameter perturbations change $G(f)$ by virtue of a slightly different interference between P, pP, and sP. These arrivals cause spectral zeroes in $G(f)$, and perturbing the mechanism causes an erratic behavior at periods less than the first spectral zero. At the longer periods the effects are less dramatic, on the order of a factor of 2 for the perturbations used. A bias of a factor of 2 in the longer period components would not affect any of the conclusions regarding characteristic rupture length.
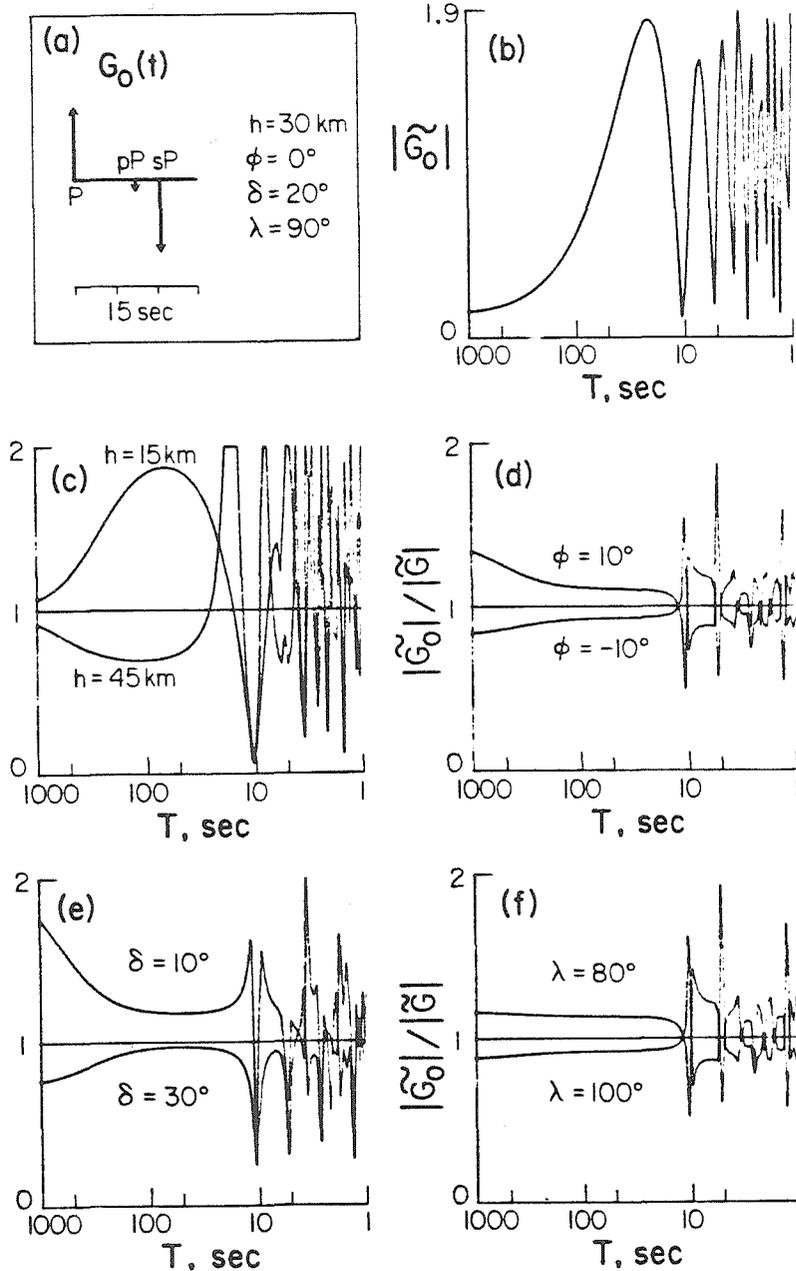
Figure 3.15. Errors in the time function introduced by incorrect focal mechanism. In (a), the half-space Green's function (no instrument or Q) is shown for a station at an azimuth of 45° and a distance of $\Delta = 90°$. The spectral amplitude of $G_o(t)$ is plotted in (b) with a linear scale (as are (c) through (f)). If our assumed Green's function differs from $G_o(t)$, then we effectively multiply the spectrum of the true time function by the functions shown in (c) through (f) for various parameter perturbations.

3.6. Source Time Functions

The time functions are calculated with focal mechanisms determined from long-period surface waves (Figure 3.3). The longer periods indicate the average mechanism over the duration of the earthquake. In most cases the first motions are consistent with the long period mechanism. Figure 3.16 shows the time functions determined for a few Niigata seismograms. The point source depth for these time functions is 10 km. Although changing the depth will change the total moment and alter the details of the time function,the basic shape and duration remain the same. The Niigata time function has a duration of 20 sec, and represents the breaking of a single asperity. The duration time can be considered as the sum of the rupture time plus the particle rise time. Assuming a rise time of a few seconds and a rupture velocity of ~2 km/sec, the characteristic length of the Niigata asperity is 30 km if the rupture is unilateral, or 60 km for bilateral or circular rupture. Abe (1975) estimated the Niigata fault area to be 30x80 km^2. Hirasawa (1965) analyzed the P-waves of the Niigata earthquake and identified two or three events, and indeed some complications can be seen in Figure 3.15. However, for the time scale of interest here, Niigata can be considered as a single event.

The inversion results for the Kurile Islands earthquake are shown in Figure 3.17. The focal mechanism from the surface wave study of Kanamori (1970a) is used, with a depth of 30 km. The most important feature is that the time function is composed of distinct events, each
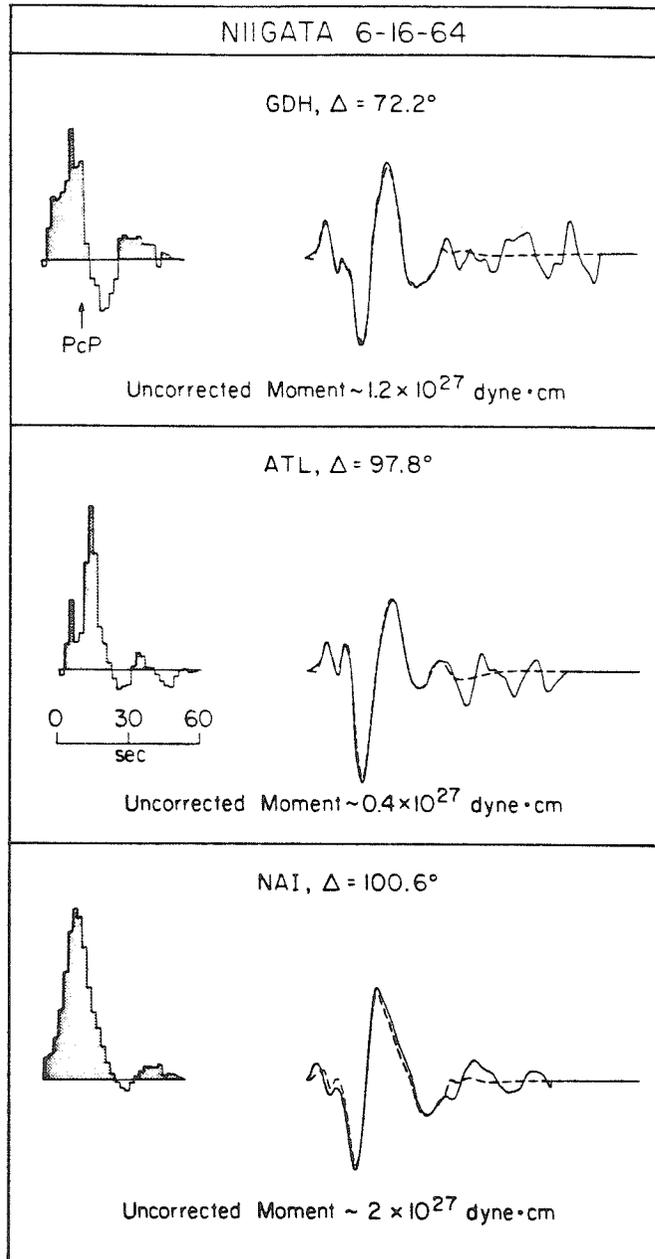
Figure 3.16. Deconvolved time functions for the Niigata earthquake. The solid traces are the observed seismograms and the dashed traces are the synthetic seismograms for the time functions shown at left. The damping constant is 10^{-2}. The common feature of these time functions is a single event with a duration of 20 sec.
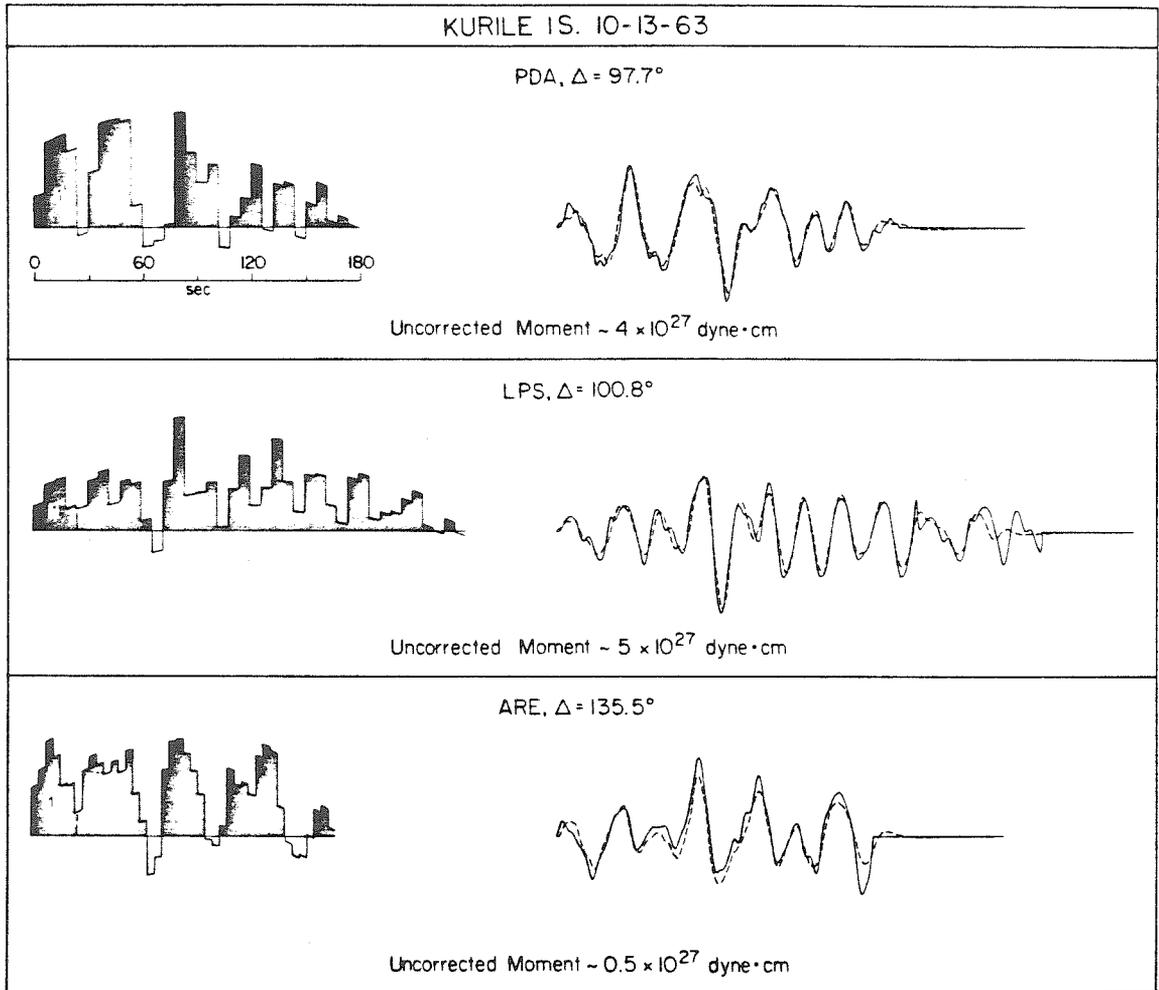
Figure 3.17. Time functions for the Kurile Islands earthquake. The value of the damping constant is 10^{-2}. The common feature of these time functions is a sequence of events with a characteristic duration of 20-30 sec and the characteristic moment of each event is comparable to that of the Niigata event.

with duration 20-30 sec. The moment of each of these pulses is approximately that of a $M_W \sim 7.5$ event. Although the total moment of these pulses does not equal the long period value of 70×10^{27} dyne-cm, this difference can be reduced by adding a long period component as previously discussed. This long period component would result from a slower deformation over the entire fault surface, while the individual pulses that dominate the P wave records are from the rupturing of asperities. The rupture of asperities is expected to cause a larger moment rate than rupture of the surrounding weaker area, due to a larger particle velocity and possibly a faster rupture rate (see equation (3.5)). Hence the breaking of asperities causes a temporal increase in the moment rate, that is the pulses seen in Figure 3.17, with the pulse duration indicative of asperity size.

"Uncorrected moment" refers to the lack of the diffraction correction to the stations in the core shadow, therefore the uncorrected moments are minimum estimates. Note that while the amplitude does decrease into the shadow, the shapes of the time functions are basically the same. It is possible to invert all of the seimograms simultaneously for a common time function. However, given the large fault areas of the earthquakes studied, we can expect real differences in the time functions due to directivity. The significance of any feature in a single station time function must be assessed by a comparison with the other single station solutions.

Figure 3.18 shows the time functions for the Rat Islands earthquake. The character of the time function is similar to that of
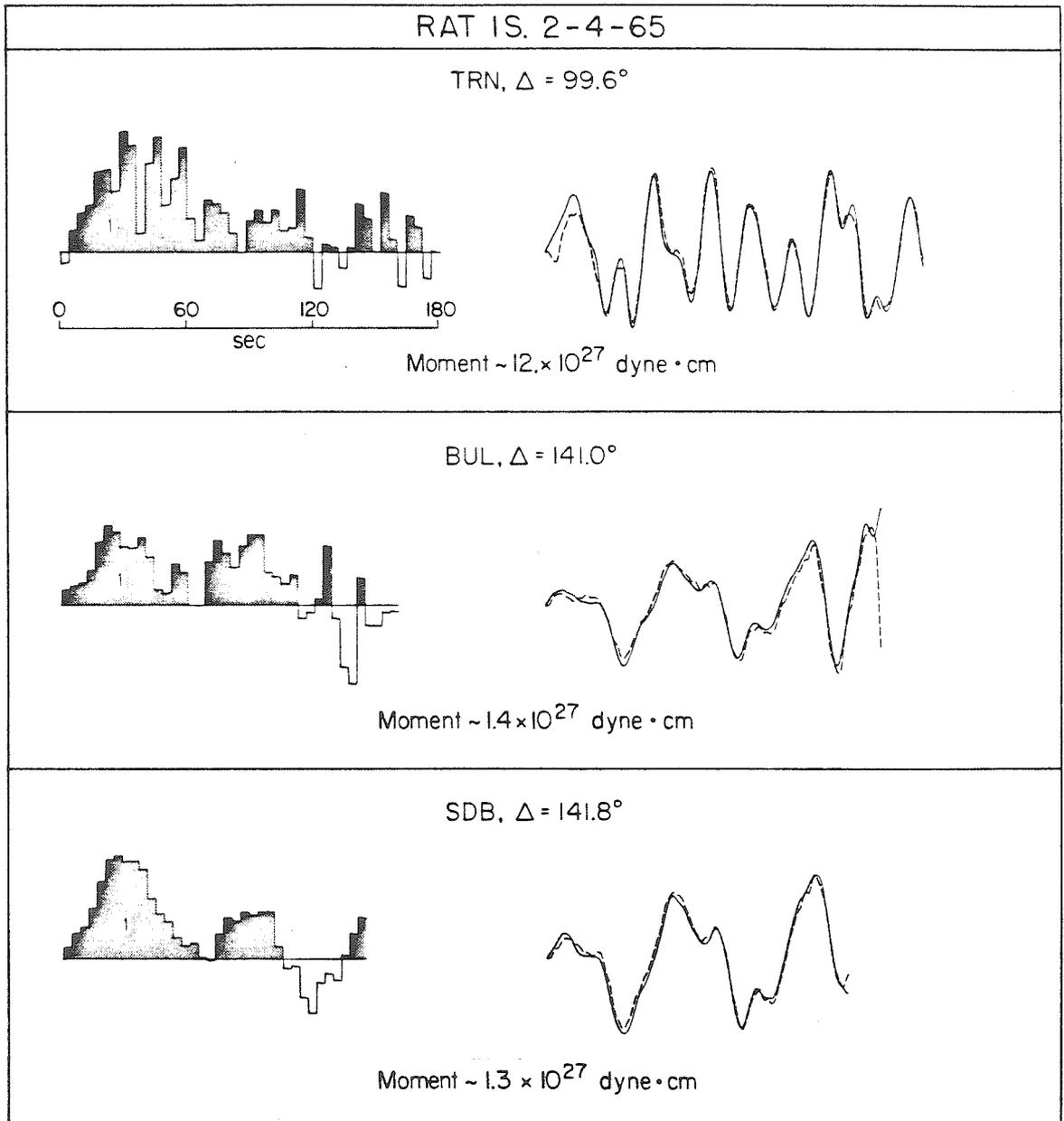
Figure 3.18. Time functions for the Rat Islands earthquakes. The damping constant is 10^{-3}. The time functions are similar to those for the Kurile Islands earthquake, except that the characteristic event duration is slightly larger and the characteristic moment is larger.

the Kurile Islands earthquake, that is a sequence of pulses. A closer examination indicates that the individual pulses have a longer duration and somewhat larger moment than the pulses of the Kurile Islands event. We will not pursue these differences however, as the Alaskan time function is strikingly different from both the Kurile Islands and Rat Islands time functions.

Alaska source time function

Figure 3.19 presents the most impressive results of the deconvolution study, though expected from our earlier discussion and the simple forward modelling. The first three minutes of the Alaskan records can be explained by one huge pulse, with a duration of at least two minutes. Recall that the forward modelling in Figure 3.10 suggested that a rising-ramp of at least 40 sec duration is required to match the first 30 sec of the Alaskan records. The inversion results indicate that a rising-ramp of 60-70 sec duration is more appropriate. A relatively smooth time function is clearly required by the data, as regardless of the damping used, the time function is invariably smooth. In fact, it seems that our time function sampling is too coarse, thereby causing a poor fit in the first part of the records. Notice the moment of this huge pulse. Although the quoted moments can easily be in error by a factor of 2, recalling that these are minimum values, we can safely conclude that the Alaskan pulse is the largest time function with the longest duration ever reported from body wave investigations.

ALASKA 3-28-64

LPB (NS), Δ = 99.3°

Moment ~ 200. × 10²⁷ dyne·cm

MUN, Δ = 120.4°

Moment ~ 45. × 10²⁷ dyne·cm

BUL, Δ = 139.0°

Moment ~ 100. × 10²⁷ dyne·cm
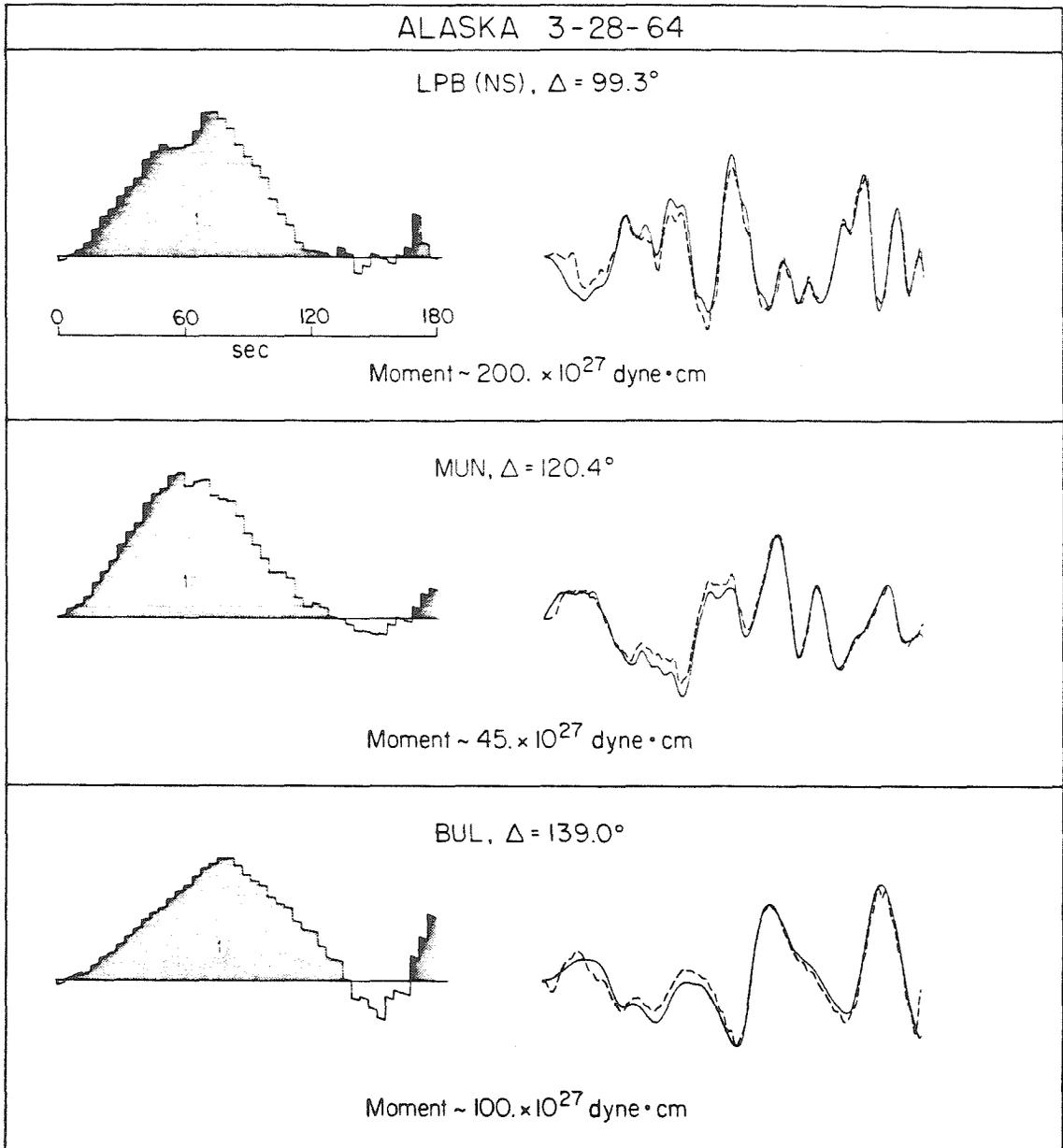
Figure 3.19. Time functions for the Alaskan earthquake. The value of the damping constant is 10^{-4}. The Alaskan time function is characterized by one huge pulse with an initial ramp of 60 sec duration. These time functions are deconvolved for the shallow source mechanism.

Before further discussion or interpretation of the Alaskan time function, it is worthwhile to examine the stability of the solution with regard to damping, focal depth, and source finiteness. Figure 3.20 shows the time functions and synthetics for moderately damped solutions, d=.01, the same as the Kurile Islands solutions in Figure 3.17. With this damping, the solutions are a poor match to the first part of the seismograms. Only by reducing the damping to d=.0001 is an adequate fit obtained. Regarding the source depth, we have used two different focal mechanisms, corresponding to two acceptable mechanisms for the long period surface waves (Kanamori, 1970b). The deep mechanism at h=70 km is inconsistent with first motions, yet inconsistent stations can be used by merely reversing the station polarity, and surprisingly obtain a compatible time function (see Figure 3.21). Using the shallow mechanism, which is consistent with first motions, the shape of the time function is essentially the same. The main difference is that the amplitude and moment is roughly twice as large. Hence, the Alaskan time function is quite stable with regard to mechanism and depth, and this is further confirmed when using a distributed source. Figure 3.22 shows the time functions for BUL calculated for the shallow fixed depth and a distributed source and the two functions are quite similar in shape and amplitude. Therefore, we conclude that there are several important features of the Alaskan time function which are required by the observed seismograms: the "first event" of the Alaskan earthquake was a huge pulse with a rising-ramp of 1 minute duration and a total duration greater than 2 minutes, and the minimum moment release of this first
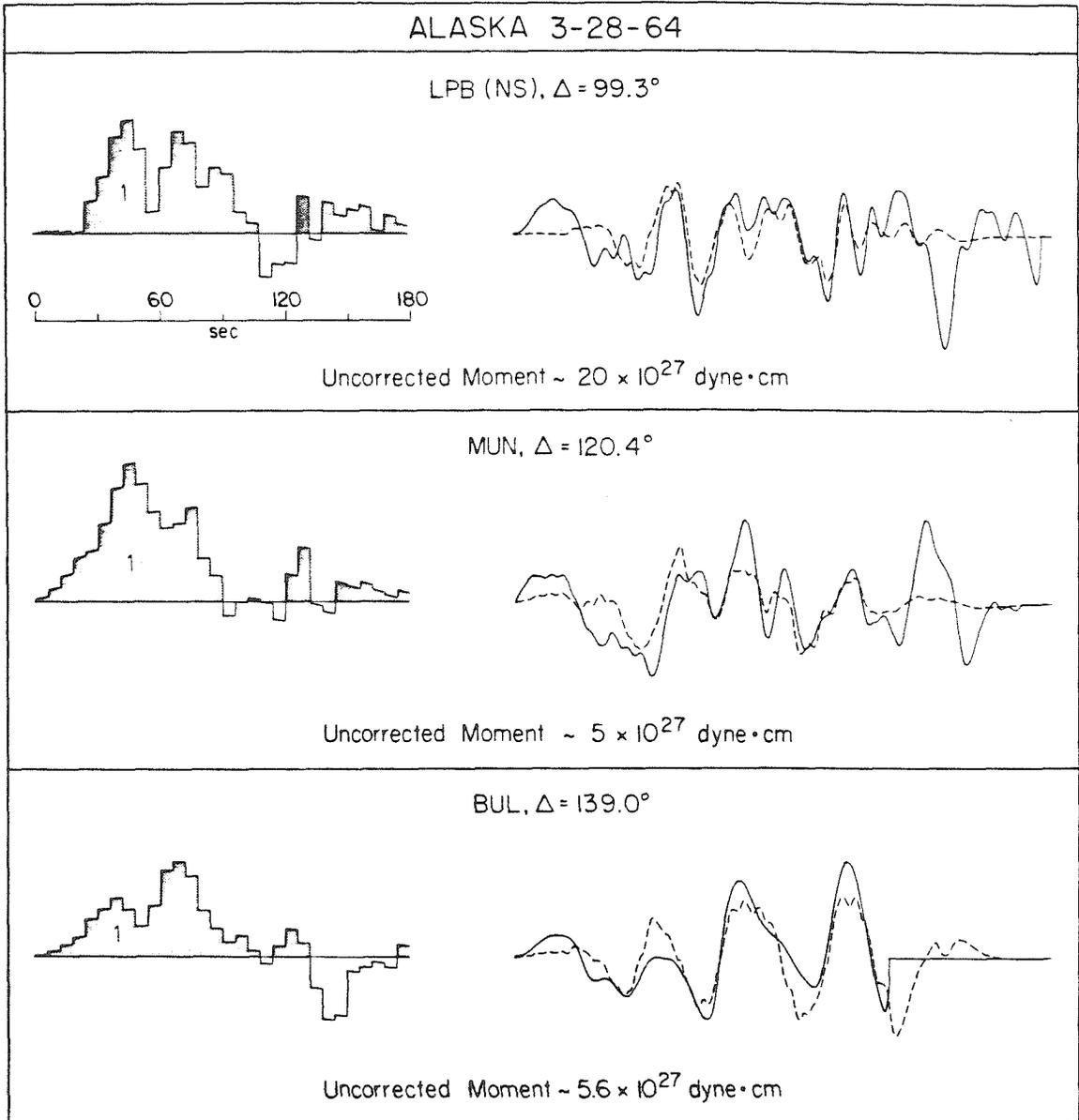
Figure 3.20. Alaskan time functions for a damping of 10^{-2}. With this damping value, the first part of the seismograms cannot be matched. The data require a substantial long period component.
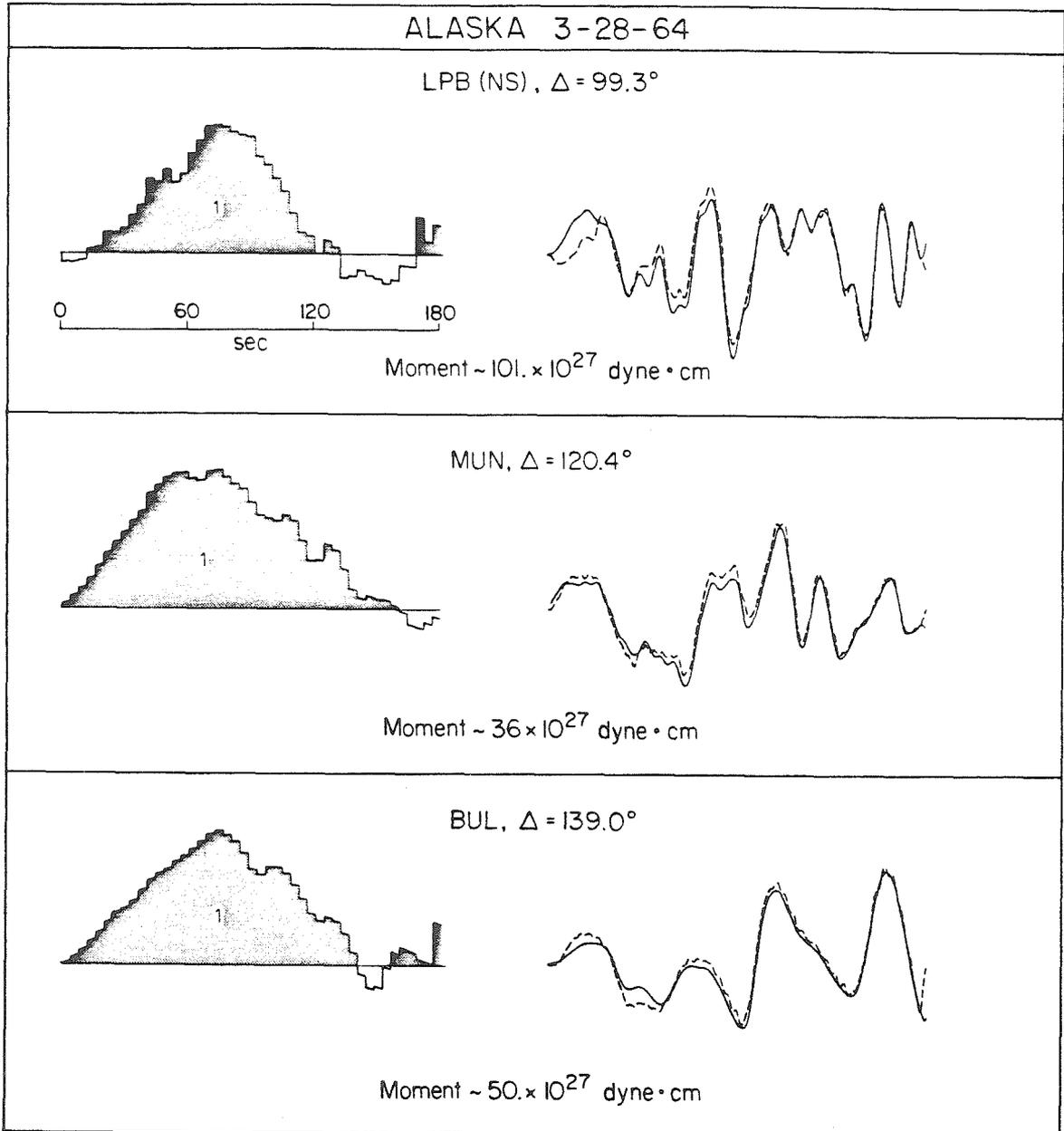
Figure 3.21. Alaskan time functions for the deep (h = 70 km) source mechanism. The damping value is 10^{-4}. The time functions are essentially the same as those in Figure 3.19 except that the moments are reduced by a factor of 2.
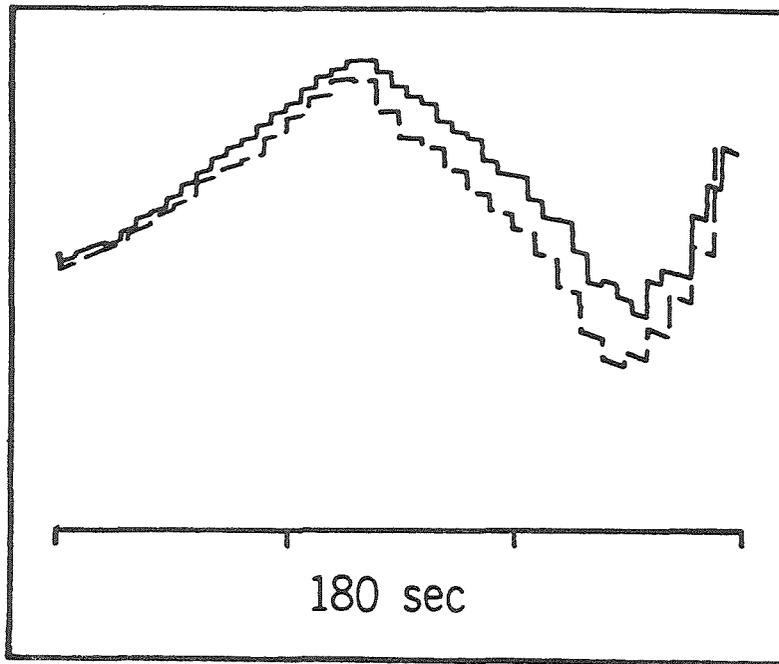
Figure 3.22. Comparison of the Alaskan source time functions at BUL for point source and distribution source. The time function in Figure 3.19 is replotted (solid). The dashed trace is the deconvolved time function for a source distributed between the surface and h = 50 km. The two functions are virtually identical in duration and moment.

pulse is 100×10^{27} dyne cm.

Source time function of the 1979 Colombia earthquake

We have included the time function of the 1979 Colombian earthquake ($M_W \sim 8.2$) in Figure 3.23. This is an interesting subduction zone event as it occurred in the region of the great earthquake of 1906 ($M_W \sim 8.8$, see Kanamori and McNally, 1981), and it also generated unusual P waves (see Figure 3.22). Other subduction zone events of this size have waveshapes similar to those of the Niigata event, though perhaps of slightly longer duration (e.g. 1969 Kurile Islands and 1976 Philippines earthquakes). The WWSSN (15-100) recordings of the Colombian event show a peculiar double pulse character. This event was well recorded by the ultra-long period (100-300) seismograph at Berkeley. Though our results are tentative, the time functions shown in Figure 3.23 are strongly suggestive of a circular type rupture, i.e. a ramp with a sharp cutoff. Notice that the duration is approximately 60 sec, thus allowing for directivity the rupture duration is at least 50 sec. Hence, for a quarter-circular rupture (based on epicentral location) and ~2 km/sec rupture velocity, the asperity is ~100 km in size. Kanamori and Given (1981) estimated the total rupture length to be 230 km, hence the asperity is approximately half the fault area. Although the moment release of the Colombian earthquake is much smaller than the Alaskan earthquake, it is interesting that the Colombian time function indicates a large asperity size within the fault area of the
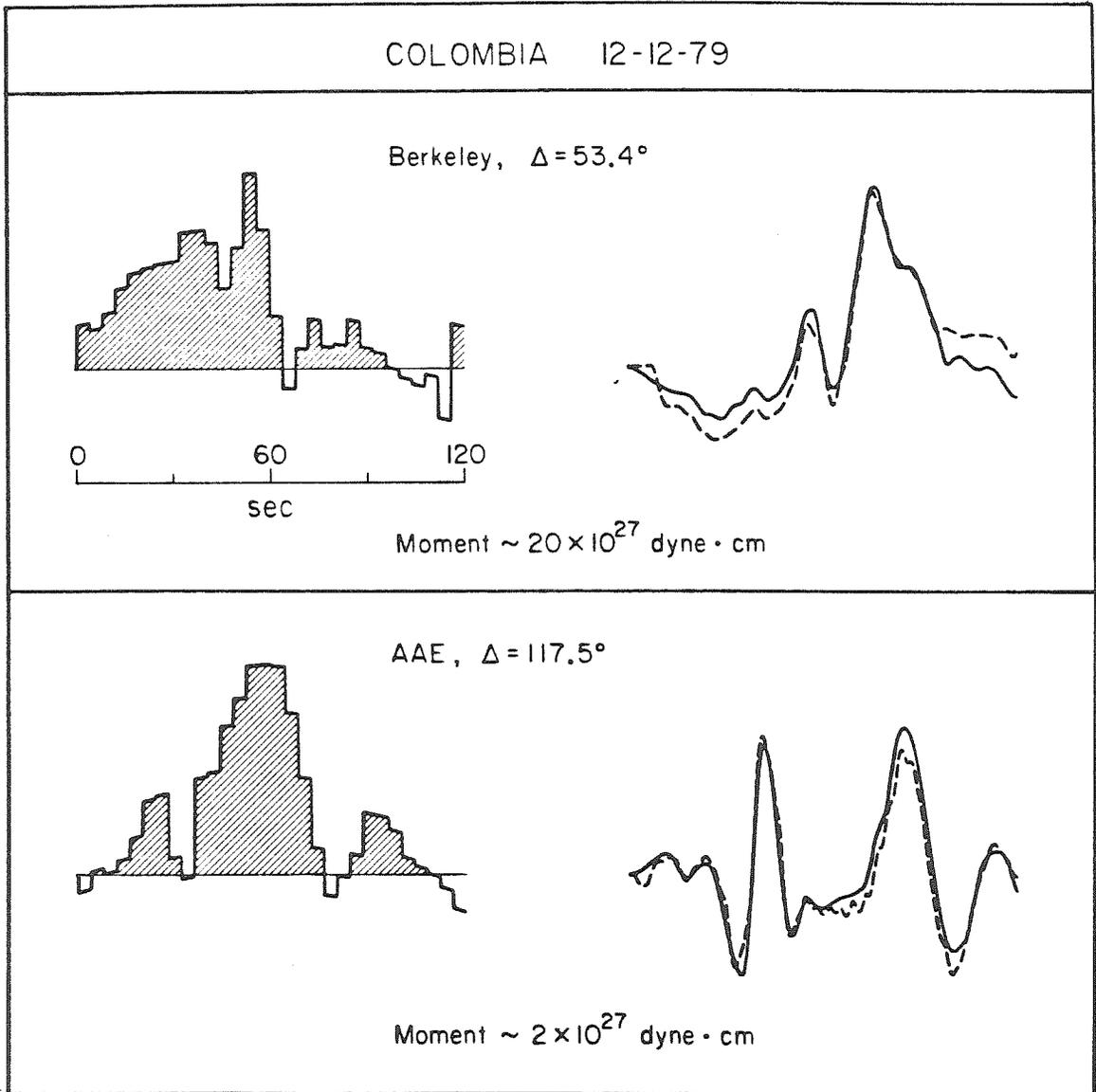
Figure 3.23. Time functions for the 1979 Colombian earthquake. The damping level is 5 x 10^{-4}. The common features of the time function at those two stations is the long duration (60 sec) and the abrupt termination at 1 minute.

great 1906 earthquake.

3.7. Interpretation

The time functions of the other events are plotted with the Alaskan time function at true relative scale in Figure 3.24. Though the Kurile Islands and Rat Islands earthquakes certainly have a long period component which would increase their amplitude, they would still appear small compared to the Alaskan time function. Of course, as the total moment of the Alaskan event is 750×10^{27} dyne cm it may not be surprising to recover a huge moment in the body waves. What may be more significant is the shape of the time function: smooth with an extremely long duration. The shape and duration of the time functions are contrasted in Figure 3.25.

Due to the unusual character and significance of the Alaskan time function, the translation of the time function into a length scale will be covered in some detail. Then, the stress drop of the Alaskan earthquake will be compared to stress drops of other subduction zone earthquakes to show that the Alaskan earthquake ruptured a strongly coupled region.

The Alaskan asperity
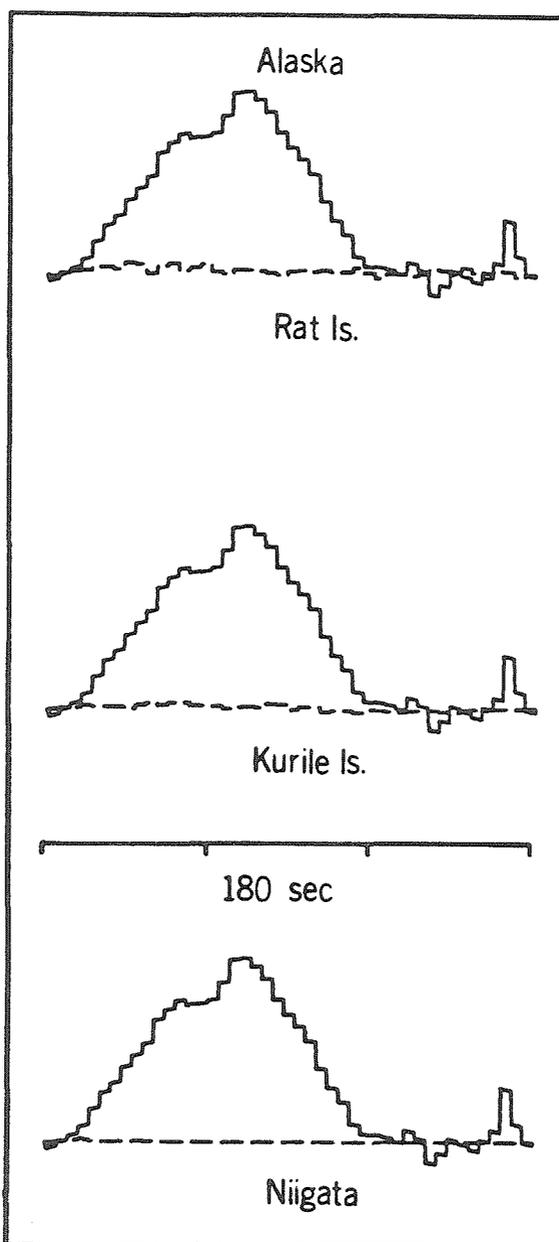
Figure 3.24. Comparison of time functions at approximately the same distance, at true relative amplitude scale. The Alaskan time function at LPB is plotted as the solid trace, while the Rat Islands (TRN), Kurile Islands (PDA), and the Niigata (ATL) time functions are plotted as dashed traces. Recalling that the amplitude scale is moment rate, a substantially higher moment rate is resolved for the Alaskan earthquake.
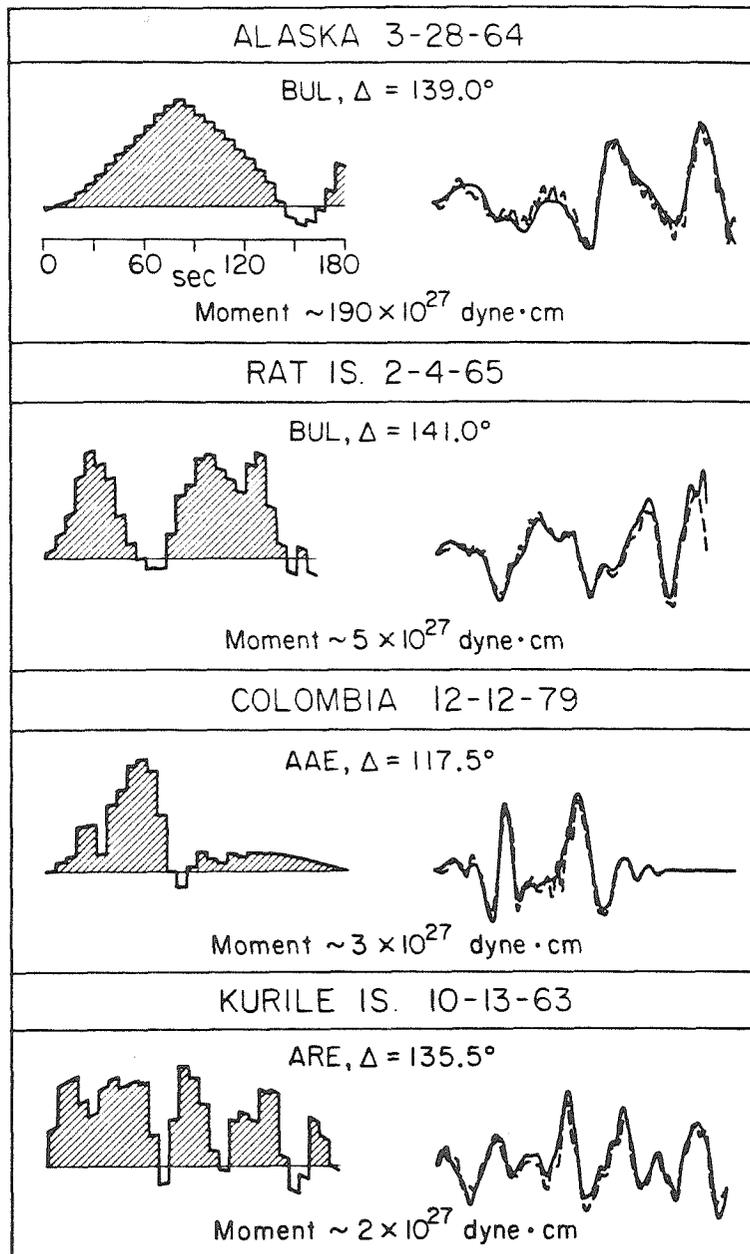
Figure 3.25. Comparison of time functions, synthetic and observed seismograms at large distance. The characteristic time scales can be directly compared, and varies from 2 minutes for the Alaskan earthquake to 30 sec for the Kurile Islands earthquake. Thus, the large size of the Alaskan earthquake is associated with a longer characteristic time, hence larger characteristic asperity length.

The time functions for the Alaskan earthquake in Figure 3.19 do not uniquely specify the rupture mode. The rupture is apparently circular-type for the first minute, but, unlike the Colombian time function which ends abruptly, the Alaskan time function continues at some level after the ramp. Therefore, the most reliable part of the Alaskan time function is the initial ramp, and we want to use this ramp to determine the length scale of rupture. The simplest approach is to assume a rupture velocity, then the radius is $R \sim V_R t_R$. A typical value of the rupture velocity is 2 km/sec, and with $t_R = 60-70$ sec the length scale is $R \sim 120-140$ km. This would be the minimum length scale, as $L \sim 2R \sim 240-280$ km if the circular-type rupture extended in all directions. We would like to estimate the length scale without simply assuming the rupture velocity. This can be done with additional information. For example, fixing the displacement allows for a L_{min} to be determined.

There are islands in the vicinity of the Alaskan epicenter which were displaced horizontally and vertically by the earthquake (Plafker, 1972; Hastie and Savage, 1970). These static deformations have been modelled by Alewine (1974) and Miyashita and Matsuura (1976) to determine the final displacements on the faults, and include imbricate faulting. Though the solutions are not unique, an average static displacement of 15-18 meters on the main fault in the epicentral region is indicated. Based on the long period surface waves, Kanamori (1970b) obtained a seismic moment of 750×10^{27} dyne-cm. Using a fault area of 15×10^4 km^2 (estimated from the aftershock distribution) and $\mu = 700$ kb, the average displacement is 7 meters (or 10 meters for $\mu = 500$

kb). Thus, the final displacement near the epicenter is twice the average. Some of the final displacement may have been post-seismic. However, using the upper bound of 15-18 meters displacement provides an estimate of the minimum scale length.

Recall from equation (3.6) that the time function results from the convolution of the particle velocity and area rate, $m(t)=\mu\dot{D}(t)*\dot{A}(t)$. If the duration of $\dot{D}(t)$ is much less than the duration of $\dot{A}(t)$, then the main effect of the convolution is to multiply $\dot{A}(t)$ by the average displacement, $m(t)=\mu D\dot{A}(t)$. To obtain the moment released up to a particular time, say τ_R, the above expression is integrated from $t=0$ to $t=\tau_R$ which gives $M_R=\mu DA(\tau_R)$. Thus the ruptured area at time τ_R is, $A(\tau_R)=M(\tau_R)/\mu D$, with minimum $M(\tau_R)$ and maximum D estimates providing a minimum estimate of the area. The Alaskan time function has a ramp time of $\tau_R=60-70$ sec, and a moment release of 10^{29} dyne-cm at this time. With $\mu=500$ kb, the fault area at τ_R is $A=1.3-1.1 \times 10^4$ km^2 for D=15-18 meters. Translating the area into a characteristic length as $L=A^{1/2}$, L=115-105 km. A unilateral rupture across a square fault of dimension 115x115 km^2 with a rupture time of 60-70 sec yields a rupture velocity of 1.7-1.6 km/sec. However, we know from the ramp shape of the time function that a unilateral rupture is not appropriate. Figure 3.26 shows three rupture modes which start as a circular-type rupture and then continue along the fault in a unilateral fashion. The location of the Alaskan epicenter relative to the aftershocks suggests that the quarter-circular rupture in Figure 3.26c is a reasonable representation of the Alaskan earthquake. In this case, $R=(4/\pi)^{1/2}A^{1/2}=130-120$ km.
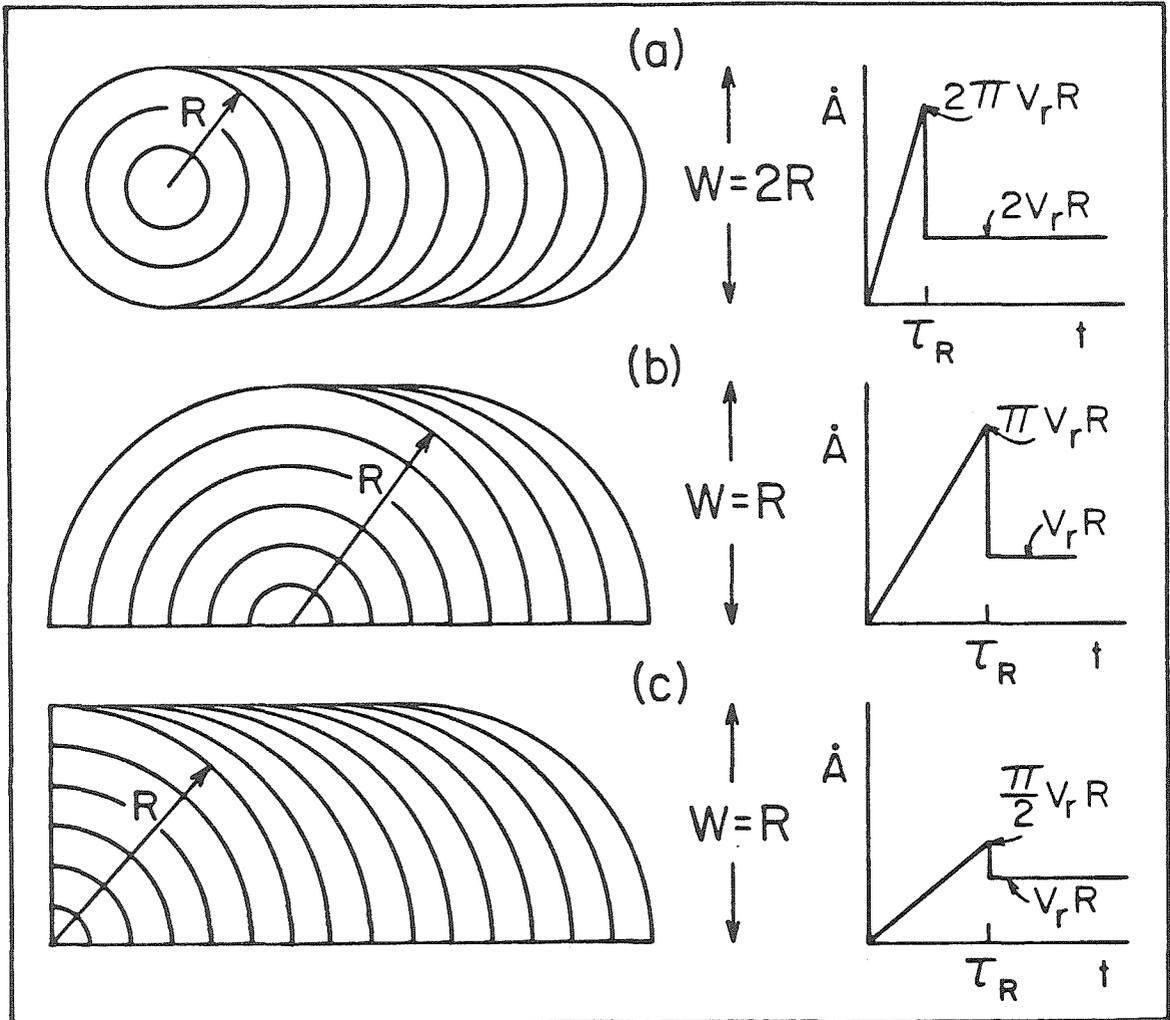
Figure 3.26. Three examples of circular-type rupture modes. The fault planes are shown at left and the progress of the rupture front from the starting point is depicted by the circular arcs. The corresponding area rates are plotted as a function of time on the right. In (a), the rupture front is circular until it expands to the fault width, at which point it proceeds unilaterally down the fault. In (b), the rupture front is semi-circular initially, while in (c) it is quarter-circular. There are many other combinations that produce an initial circular-type rupture and thus a ramp time function.

With τ_R=65 sec, we obtain a rupture velocity of 2.0-1.8 km/sec. Therefore, in assuming a seismic displacement of 18 meters over the entire northeastern segment of the fault, we find that the minimum scale length ruptured in the first minute is 120 km. For a reasonable upper bound on the scale length, if the seismic displacement is just 10 meters and $M(\tau_R)$= 2×10^{29} dyne-cm, then R=240 km. In support of this length scale, Kanamori (1970b) suggested that the fault width is at least 200 km on the basis of both the aftershock pattern and surface wave directivity. The distribution of aftershocks is quite interesting, as there seems to be a concentration of aftershocks within 200km of the epicenter (Figure 3.27). This region may be a huge asperity.

Regarding the initial rupture length scale, the results of Wyss and Brune (1967) are of interest. Wyss and Brune used short period records to identify multiple events. In view of the small short period amplitudes relative to the long period amplitudes, it appears that the short period multiple events are small glitches on the rising-ramp time function. However, a well correlated short period pulse could serve as a "marker". The largest and most dependable short period pulse (event C of Wyss and Brune) is located at the edge of the concentration of aftershocks (Figure 3.27) and is delayed by 44 seconds. This supplies additional evidence that the initial rupture length scale extended to approximately 200km in the southwest direction within the first minute of rupture.
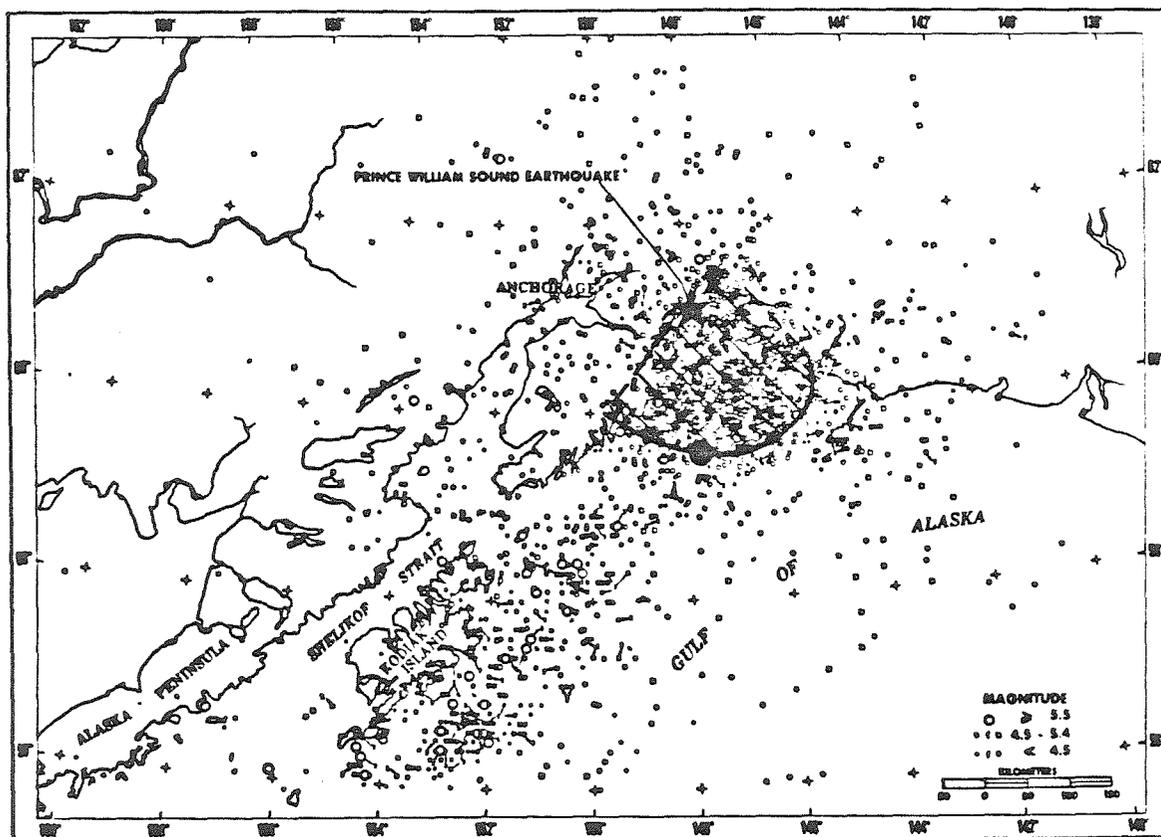
Figure 3.27. The Alaskan asperity. The aftershocks of the 1964 Alaskan earthquake (epicenter indicated by star) are plotted for a 20 month period (from Algermisson et al., 1969). The small open squares and circles are aftershocks located with the temporary local network. Notice the concentration of aftershocks in the northeastern part of the fault zone (the rupture extended to south of Kodiak Island). The heavy line, which delimits the high aftershock activity, is at 140 to 180 km from the epicenter. This distance is quite comparable to the length scale of the initial circular-type rupture estimated from the time function. Also, the location of event C (plotted as the large dot) from the study of Wyss and Brune (1970) provides additional evidence that the ruptured area extended to 180 km in the first minute. We conclude that the hachured region is a giant asperity.

Stress drops

The asperity model of earthquake occurrence stipulates that the stress is gradually concentrated at the largest asperities. If the stress level returns to a constant value on the fault plane after the earthquake, then the stress drop at the asperity should be larger than the average stress drop.

If the entire Alaskan fault plane ruptured homogeneously, then the moment release should be uniform over the fault plane. Using the static moment relation, $M=\mu DA$, the average displacement over the entire fault is 10 meters (with 7.5×10^{29} dyne-cm and 1.5×10^5 km^2, Kanamori, 1970b, and $\mu=$ 500 kb, the value used for all subseqent calcualations). If the average displacement for the first minute of rupture is 15 meters, then the average displacement over the remainder of the fault is 9 meters. Hence, the moment release could be concentrated by nearly a factor of 2 in the northeastern "asperity".

Stress drop is a poorly determined seismological parameter. The fault dimension is usually estimated on the basis of the aftershock area, and the displacement is usually deduced from the fault dimension and far-field measurements of the seismic moment. These quantities can be in error by a factor of 2. Hence, it is with some reluctance that we quote values of stress drop and no arguments below the factor of 2 level will be presented.

The stress drop is $\Delta\sigma=c\mu D/L$, where D is a characteristic displacement, L is a characteristic fault length, μ is the local shear

modulus, and c is a geometric constant of order 1. Examples of the value of c are: for a strike-slip fault of infinite length, taking L to be fault width, c=0.64, for a dip-slip fault of infinite length, with L as the fault width and a Poisson's ratio of 0.25, c=0.85, and for a buried circular fault with diameter L, c=2.7, or with $L=A^{1/2}$, c=2.4. The value of c varies for finite near-surface faults with different dips (Sykes and Quittmeyer, 1981). The characteristic length L is frequently taken as the square root of fault area. This might cause some bias if a fault is long and narrow. In the following presentation, we will write the stress drops with c as an undetermined constant. With $L=A^{1/2}$ and $D=M/\mu A$, the stress drop relation is $\Delta\sigma=cM/A^{3/2}$. The average stress drop for the Alaskan rupture is then $\Delta\sigma=c13$ bars. If the fault area is 12×10^4 km^2 instead of 15×10^4 km^2, then $\Delta\sigma=c18$ bars.

To calculate the stress drop of the northeastern asperity, the stress drop can be recast as $\Delta\sigma=c(\mu D)^{3/2}/M^{1/2}$. Assuming that we can apply this formula to the rupture as a function of time, the stress drop of the first minute of the Alaskan rupture is $\Delta\sigma=c65$ bars ($M=10^{29}$ dyne cm, $\mu=500$ kb, and D=15 m). If the moment is doubled to 2×10^{29} dyne-cm, then $\Delta\sigma=c46$ bars. To produce a stress drop of c20 bars, the displacement would have to be 10 meters or less. Thus, we tentatively conclude that the stress drop in the first minute of rupture is two to three times the average stress drop.

If one does not want to accept this, then we can return to the assumption concerning the seismic displacement in the epicentral region. Instead of choosing the upper bound, one could assume that the

displacement in the first minute was less, say 8 meters, and that the displacement continued as the rupture front moved into other portions of the fault. In essence, the duration of $\dot{D}(t)$ is no longer small compared to the duration of $\dot{A}(t)$. This possibility cannot be ruled out, and indeed from numerical studies there is some indication that $\dot{D}(t)$ has a longer duration for a larger fault area (e.g. Mikumo and Miyatake, 1978). However, if the eventual seismic displacement in the northeastern segment is 15 meters and the displacement elsewhere is only 9-10 meters, then we see that the final strain change, hence stress drop, is still larger in the epicentral region, though we cannot calculate its value so simply. Therefore, under the assumption that the seismic displacement in the northeastern segment was 15 meters or more, the stress release was greater in this portion, hence a giant asperity.

The smooth rupture of the Alaskan earthquake might cause one to think that the entire region is relatively weak. A comparison of average stress drops shows that the Alaskan fault zone was at least as strong as any other large earthquake region. For the 1965 Rat Islands event, using $M=125 \times 10^{27}$ dyne-cm and $A=7.8 \times 10^4$ km^2, we have $\Delta\sigma=c6$ bars. The 1963 Kurile Islands Earthquake has an average stress drop of $\Delta\sigma=c7$ bars for $M=67 \times 10^{27}$ dyne-cm and $A=4.4 \times 10^4$ km^2 (Kanamori and Anderson, 1975). To compare the Alaskan earthquake to events in which just one large asperity has broken, the average stress drops of recent large subduction zone events (all except for Niigata) are listed in Table 3.2. These earthquakes are considered as examples of just one large asperity failing, unlike the Kurile Islands and Rat Islands earthquakes. It

TABLE 3.2

Earthquake	M_W	M_o (x10^{27} dyne-cm)	A(x10^3km^2)	$\Delta\sigma$(bars)
1964 Alaska	9.2	750[2]	150[2]	c13
(asperity)		100	13	c65
1965 Rat Islands	8.7	125[1]	78[1]	c 6
1963 Kurile Islands	8.5	67[1]	44[1]	c 7
1979 Colombia	8.2	29[3]	17[3]	c13
(asperity)		20	8	c28
1968 Tokachi-Oki	8.2	28[1]	15[1]	c15
1969 Kurile Islands	8.2	22[1]	15[1]	c12
1966 Peru	8.1	20[1]	11[1]	c17
1976 Philippine	8.1	19[4]	13[4]	c13
1964 Niigata	7.6	3.2[5]	2.4[5]	c27
1978 Oaxaca	7.6	3.2[6]	5.5[6]	c 8
(asperity)		3.2	2.8	c22

(1) Kanamori and Anderson, 1975.

(2) Kanamori, 1970b.

(3) Kanamori and Given, 1981.

(4) Stewart and Cohn, 1979.

(5) Abe, 1975.

(6) Stewart et al., 1981.

seems that the Alaskan rupture was not a weak zone with a substantially lower stress drop than "single asperity" earthquakes. There are two earthquakes, Oaxaca and Colombia, in which a guess can be made of the asperity stress drop, and these stress drops are comparable to the stress drop estimate of the giant Alaskan asperity.

3.8. Conclusions

The main conclusion of this chapter is that the characteristic asperity scale length increases as M_W increases, with the extreme of large scale length represented by the Alaskan earthquake. This conclusion stems directly from the large differences in the observed seismograms.

The seismograms are modelled to determine reliable quantitative features of the time functions. The Kurile Islands earthquake is characterized by a sequence of pulses, each pulse of 20-30 sec duration with a seismic moment $<10^{28}$ dyne cm. Thus, we can view this rupture process as a sequence of magnitude 8 events plus an overall longer period component. Each pulse represents the breaking of a single asperity with a scale length of 40-60 km. The Rat Islands earthquake is similar in character to the Kurile Islands earthquake, except that the moment of the individual pulses is larger and the duration is somewhat longer, indicating an asperity scale length of 40-80 km. The 1979 Colombian earthquake began as a circular-type rupture with a single dominant pulse of 1 minute duration. The associated length scale is

100-120 km.

The Alaskan earthquake is characterized by a smooth rupture for at least three minutes, the initial rupture consisting of a circular-type rupture for at least the first minute. The relative lack of energy in the shorter periods allows us to directly observe the longer period components of the time function (e.g. Figure 3.8). It seems that the length scale associated with the first minute of rupture is 140-200 km. Although the entire Alaskan earthquake may have been a relatively smooth rupture, the northeastern (epicentral) segment is a giant asperity as the moment release was greater in this region, if the observed static deformation was co-seismic. The average stress drop of the Alaskan earthquake (c13-18 bars) is no smaller than the average stress drops of other subduction zone earthquakes (ranging from c6 to c17 bars). Using information from the static deformation, the stress drop estimated for the giant asperity is c45-65 bars.

We conclude that the Alaskan earthquake occurred due to the failure of a smooth, strong asperity with a dominant length scale no less than 140 km. The rupture process of great earthquakes varies from a sequence of magnitude 8 events (e.g. the Kurile Islands earthquake) to the limiting case of a broad contact surface that is coupled by a giant strong asperity (e.g. the Alaskan earthquake).

Chapter 4

A Simple Model of Seismic Coupling and

Uncoupling at Subduction Zones

4.1. Introduction

In Chapter 2, we found that earthquake size (i.e. M_W) is correlated with two other parameters: age of the subducting lithosphere and convergence rate. The largest earthquakes occur in subduction zones with a fast convergence rate and a young lithosphere, while the relatively aseismic subduction zones are those with a slow convergence rate and old lithosphere. Additionally, the penetration depth and horizontal extent of the Benioff zones are primarily correlated with the lithosphere age and convergence rate respectively. The deep Benioff zones are associated with old lithosphere and, a large horizontal extent is associated with a fast convergence rate. These latter two correlations imply that the geometry of the Benioff zone is strongly influenced by the sinking velocity of the slab as determined by the age (older lithosphere is presumably denser) and the horizontal velocity of the slab is given by the convergence velocity. These correlations can be combined into the simple concept of "preferred trajectory", where the subducting slab prefers to descend in a particular direction dependent upon convergence rate and lithosphere age. Thus, preferred trajectory parameterizes the correlation of earthquake size to age and rate. The largest earthquakes occur in zones where the preferred trajectory is more horizontal, and the aseismic zones are those in which the preferred trajectory is more vertical. Certainly, the simple parameter of preferred trajectory can only explain the general trend. To enable a

physical connection between preferred trajectory and earthquake size, we need to understand how the size of earthquakes is related to the mechanical conditions between the two plates.

The rupture processes of three great earthquakes were studied in Chapter 3. The considerable range in M_W for these great earthquakes allows a test of whether variation in earthquake size is reflected by a variation in rupture process. There is in fact quite a difference in the rupture history of these events, which can be interpreted in terms of an asperity model of earthquake occurrence. The asperity model assumes that the fault plane is composed of weak and strong regions. As shear stress is applied to the fault, the weak regions slip (either aseismically or as small earthquakes), thereby loading the strong regions (asperities). Eventually only the largest and strongest asperities are unbroken. These asperities then break as the largest earthquakes in that region. (see Lay and Kanamori, 1980; Kanamori, 1981; Lay and Kanamori, 1981; Rudnicki and Kanamori, 1981). The results of Chapter 3 support the idea that asperity size at least partially controls earthquake size. The Kurile Islands earthquake is characterized by the successive breaking of asperities, each one of "typical magnitude 7.5" size, i.e. 50 km length scale. In contrast, the Alaskan earthquake is characterized by the smooth rupture of an extremely large asperity of length scale 140-200 km. Figure 4.1 summarizes the interpretation of the time functions and length scales for the Alaskan and Kurile Islands earthquakes. Thus, even among the great earthquakes, we see that there is a difference in the mechanical
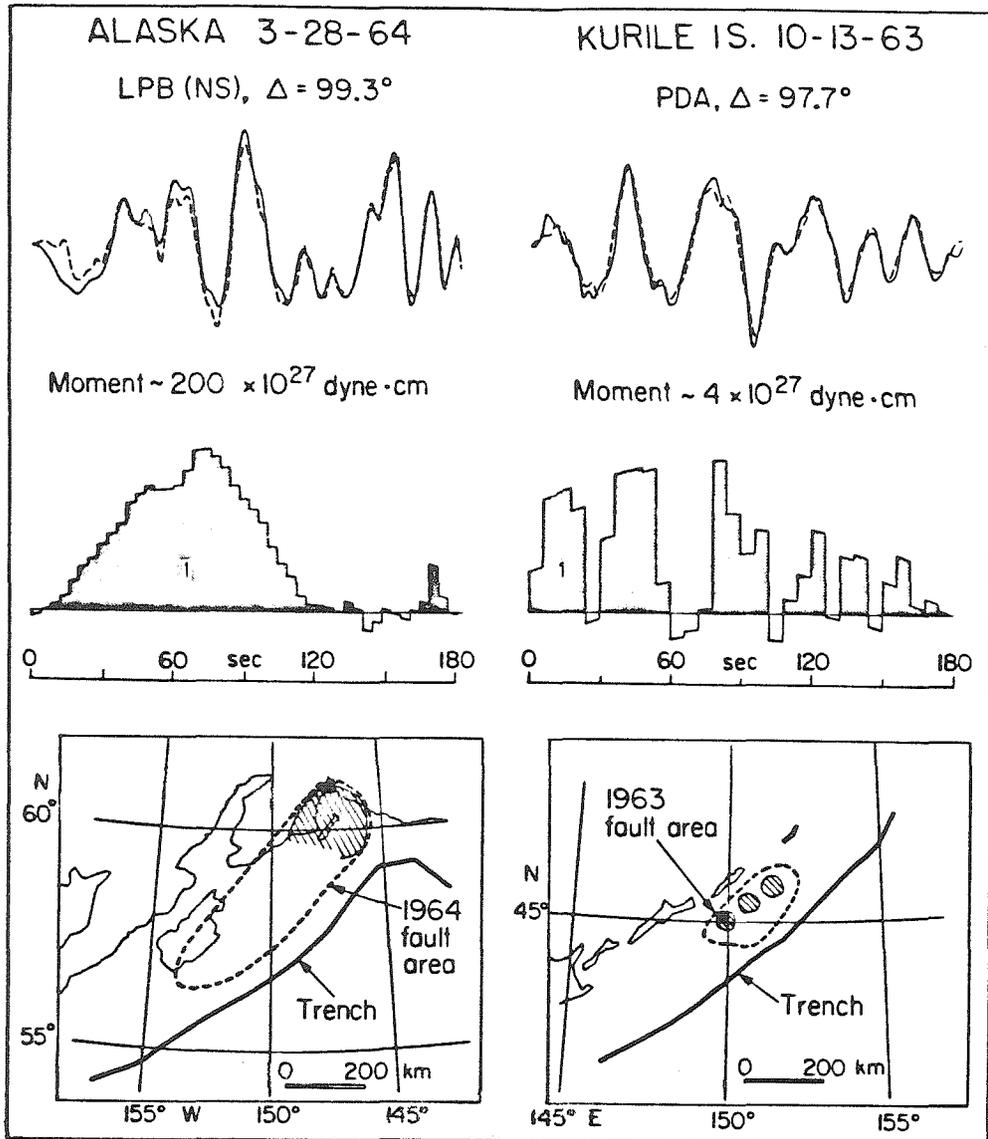
Figure 4.1. The difference in the source time histories and inferred asperity distributions of the 1964 Alaskan and 1963 Kurile Islands earthquakes. The seismograms are shown at top as the solid traces. The dashed traces are the synthetic seismograms corresponding to the deconvolved source time functions shown beneath the seismograms. While the Kurile Islands time function is composed of a multiple event sequence, the Alaskan time function represents a smooth rupture spreading over a large area. Also, notice the large difference in the seismic moments, i.e. scale of the time functions. The epicenters (stars) and fault areas are shown below, and the inferred asperities are indicated as the hachured areas. The Alaskan earthquake occurred in a region of very large asperity length scale.

condition of the contact zone, with the largest earthquake characterized
by a smooth, strong, contact surface.

4.2. Asperities and Seismic Coupling

The time functions determined in Chapter 3 present a consistent
pattern with a larger M_W associated with larger characteristic asperity
size. The time function of the 1979 Colombian earthquake also follows
this trend if we characterize that subduction zone by the great
earthquake (M_W=8.8) in 1906. We assume that this pattern applies to
subduction zones in general. We need to review this assumption in the
future as more information becomes available.

If we accept that seismic coupling is related to asperity size,
then we can develop a simple model as shown in Figure 4.2. The basic
idea is that the average shear stress (σ_t) across the fault zone is
borne by the asperities. Hence, the stress is concentrated at the
asperities, with the asperity stress (σ_a) proportional to the total
fault area (S) divided by the summed asperity area (Σs_a), that is, $\sigma_a \sim$
$\sigma_t(S/\Sigma s_a)$. In the two-dimensional cross-section of Figure 4.2, the
asperity stress would be $\sigma_a \sim \sigma_t(L/\ell)$. This simple two-dimensional
model ignores the variation in asperity size along the strike of the
subduction zone, although the lateral dimension may be the minimum
length in some cases. Based on stress drop considerations, we can assume
that the breaking stress of asperities is approximately the same
regardless of size. Then the characteristic maximum shear stress in the
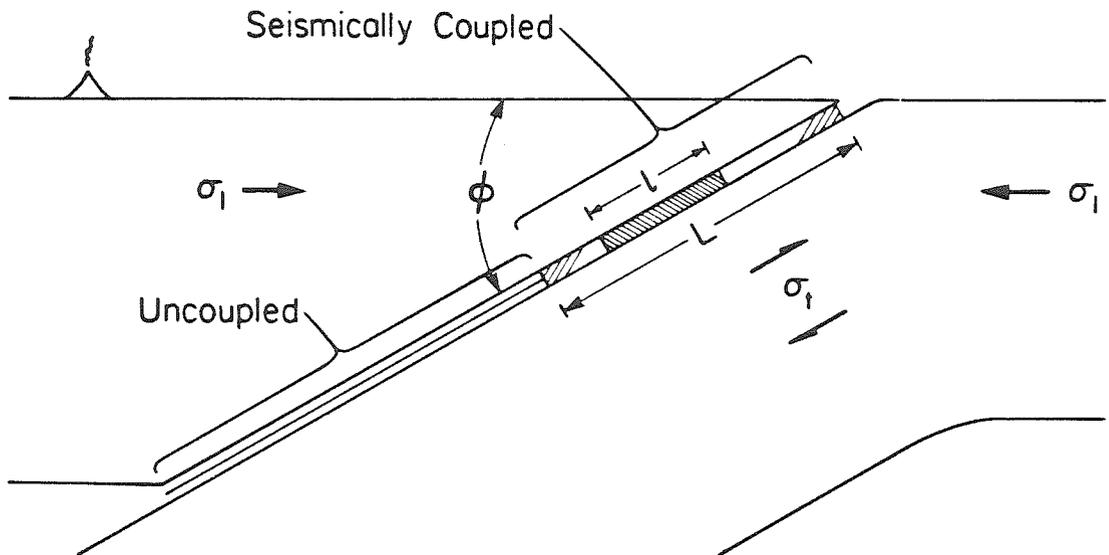
Figure 4.2. A simple model of seismic coupling. The seismically coupled region extends down to ~40 km depth. All subduction zones appear to be uncoupled below this depth. The tectonic shear stress (σ_t) is transmitted across the contact zone (L) mostly by the asperities, where we have shown one of length ℓ. The horizontal compressive stress (σ_1) is proportional to σ_t. The size of the asperities is quite variable, i.e. ℓ can extend to the maximum length L (e.g. Alaska), while some zones have a very small characteristic ℓ, essentially completely uncoupled (e.g. Marianas). As discussed in the text, a larger ℓ corresponds to a higher average σ_t and σ_1.

plate is: $\sigma_t \sim \sigma_b(\Sigma s_a/S)$, with σ_b the breaking stress. At this level of σ_t the asperity(ies) will be loaded to the breaking stress, thereby breaking the asperity(ies). The average "tectonic" stress, σ_t, can be resolved into a horizontal compressive stress, σ_1 in Figure 4.2, as $\sigma_1 \sim$ $\sigma_t/(\sin\phi\cos\phi)$. All of these relations are only approximate, but they serve to indicate how we can translate the size of earthquakes, via characteristic asperity size, into "tectonic stress" in the slab, and consequently overall horizontal compressive stress between the two plates. Notice that the stress drop measured for different size earthquakes could be the same, but that the average shear stress in the lithosphere will be larger for larger earthquakes. Thus, this simple model of seismic coupling relates earthquake size to the total stress between the plates, and hence agrees quite satisfactorily with the intuitive idea of seismic coupling that has emerged in the previous papers on the subject.

An interesting subject is the interaction of displacement in the coupled and uncoupled regions. One possibility is that a creep event in the deep uncoupled region might load the shallow coupled region, consequently causing an earthquake. Kanamori and Cipar (1974) reported a slow precursor to the 1960 Chile earthquake. This precursor might have occurred in the uncoupled region, which is normally aseismic. Another aspect is that if slab pull is generally important in loading the seismically coupled region, then the stress would tend to be concentrated at the deepest portion of the coupled region. We might then expect the rupture of large earthquakes to start at the bottom of

the coupled region and to extend primarily updip and along the subduction zone. The 1964 Alaska (see Figure 4.1) and 1965 Rat Islands (see Wu and Kanamori, 1973) great earthquakes are just two examples of this. In fact, there appears to be some evidence for the concentration of stress at the base of the coupled zone in the Shumagin (Alaska) region (House and Boatwright, 1980; also Sykes, 1981, personal communication). Thus, understanding the interaction between the shallow seismically coupled region and the deeper uncoupled region of the contact zone may yield information on the earthquake cycle.

To briefly state this model of seismic coupling; a stronger coupling corresponds to a larger proportion of asperity area, thereby a higher average shear stress in the two plates in the depth range 0 to ~40 km, and consequently a higher tectonic horizontal compressive stress. There are two questions that immediately appear: (i) why are the plates seismically uncoupled below ~40 km (as the overlying lithosphere extends down to ~100 km),and (ii) what are the asperities and what causes the seemingly systematic variation in their size scale. We will consider these questions in the next two sections.

4.3. Plate uncoupling at depth

Large subduction zone earthquakes apparently do not rupture much deeper than ~40 km (based on estimates of fault extent from aftershocks, see Kelleher et al., 1974; Kanamori and Anderson, 1975 for references to specific studies, also Sykes and Quittmeyer, 1981). Though there is

seismicity below 40 km, the earthquakes are substantially smaller than the large shallow events, and many of these events occur entirely within the subducting lithosphere. Thus, even in subduction zones with great earthquakes, the plates seem to be uncoupled below 40 km. There must be a significant change in the properties of the contact zone. One possibility is that the overlying lithosphere can be easily deformed below 40 km (i.e. the mechanical lithosphere of the overlying plate is only 40 km thick). Or, if there is a distinct material in the contact zone at this depth, perhaps a rheological change occurs in this material at 40 km that allows low-stress creep between the two plates. However, the possibility that we pursue here is a change in the down-going slab.

The composition of the oceanic crust is basaltic according to the current consensus. Evidence for this includes: extruded rocks at the mid-ocean ridges, rocks dredged and drilled in the oceanic basins, and seismic refraction profiles and density models. (We will use basalt as a compositional classification, thereby including gabbro). At high pressure, basalt transforms to eclogite with a density change of approximately 15% (Anderson, 1979; also see Wyllie, 1971, for a general discussion of the basalt to eclogite phase change). Assuming that the oceanic crust is subducted along with the oceanic lithosphere, the basalt to eclogite transformation should occur in some depth interval if the reaction rate is fast enough. In this depth interval, the phase transformation could cause a drastic reduction in the shear strength of the rock through the mechanism of superplasticity (to be described later). If so, then the 6 km thick crust might serve as an

uncoupling zone between the two plates. For this mechanism to operate,
two processes need to occur: (i) the basalt to eclogite phase change,
and (ii) that this phase change be accompanied by superplastic behavior.
As will be discussed below, it seems reasonable that these two processes
will occur.

Basalt→Eclogite

The basalt→eclogite phase change has been investigated
experimentally (e.g. Yoder and Tilley, 1962; Cohen et al., 1967;
Ringwood and Green, 1966; Green and Ringwood, 1967) prompted by the
hypothesis that the Moho discontinuity is a phase change of the basaltic
crust to eclogite. Most of these tests were done under anhydrous
conditions and with compositions different from oceanic tholeiites.
Also, the experiments were usually run at slightly sub-solidus P-T
conditions (see Figure 4.3). Calculations of the subduction zone
thermal regime (e.g. Hsui and Toksoz, 1979) indicate that the top of
the subducting slab should be exposed to temperatures of $200°-500°C$. To
determine the pressure interval over which the phase change occurs at
these low temperatures, the experimental results have to be extrapolated
to lower temperatures. However, the data in Figure 4.3 do not allow a
reliable extrapolation. Still, the experiments are useful as they
provide an upper bound on the transformation pressure. If we assume
that $dP/dT=0$, then the phase change should start to occur no deeper than
53 km (Cohen et al.,1967) or 40 km (Green and Ringwood, 1967). Also,
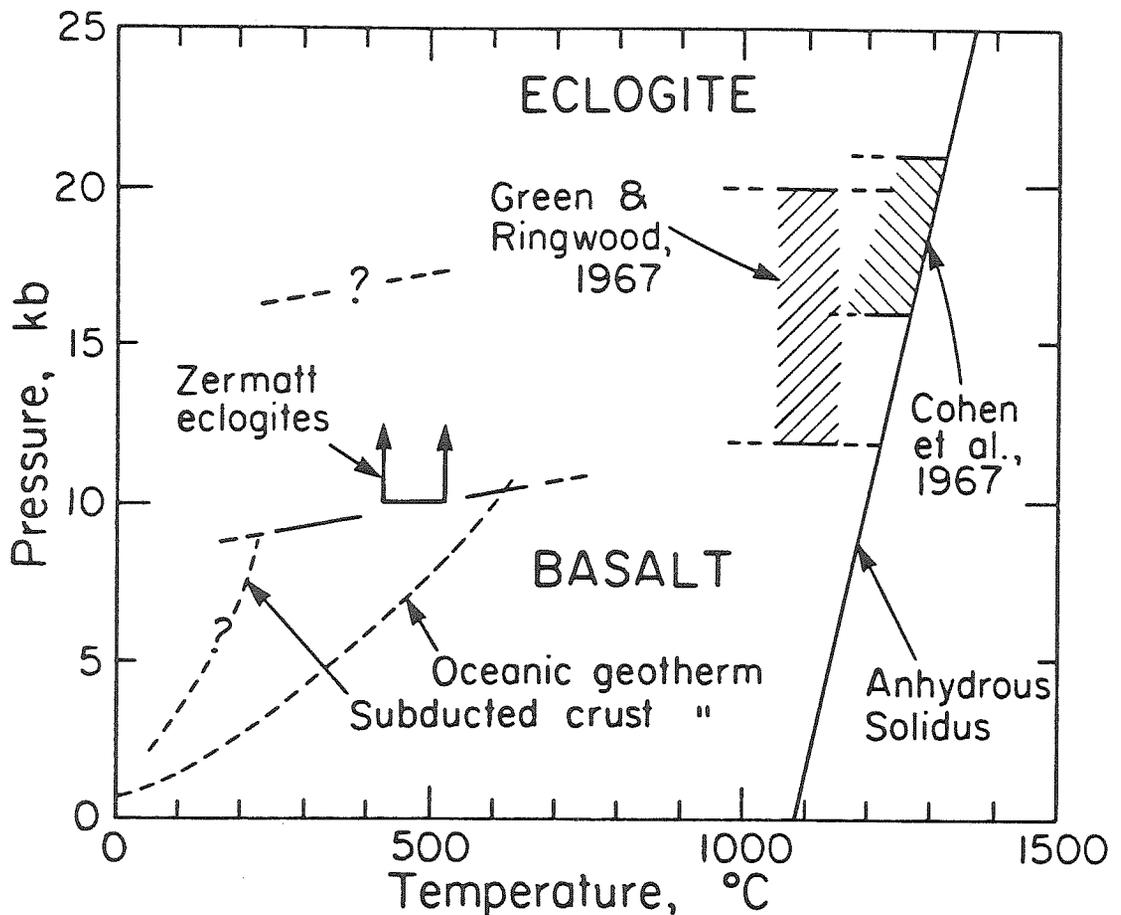
Figure 4.3. The basalt to eclogite transformation. Experimental determinations of the anhydrous solidus and transition intervals are plotted (stipled regions). Cohen et al. used an olivine tholeiite, while the Green & Ringwood results for an alkali olivine basalt are shown to demonstrate the dependence on composition. The Zermatt eclogites (representing subducted oceanic crust, see Ernst, 1981) indicate that the basalt to eclogite transition would start at 10 kb (30-35 km depth) under conditions close to those expected in the subduction zone. The upper boundary of the transition is uncertain.

the transformation pressure interval of the experiments is 5-7 kb, corresponding to a depth interval of 17-23 km. Regarding the reaction rate, a theoretical study by Ahrens and Schubert (1975) concluded that a very small amount of water would lead to a rather fast reaction from basalt to eclogite, and we can expect the oceanic crust to be wet throughout (e.g. Lister, 1977).

The experimental and theoretical studies indicate that oceanic basalt should transform to eclogite, starting no deeper than 40-50 km. There are field observations showing that oceanic basalt does in fact transform to eclogite. An excellent example is in the Zermatt region of the Alps, where recognizable pillow basalts have been at least partially converted to eclogite (Bearth, 1959). In addition, the regional setting of these eclogites suggests that they formed in a subduction zone environment, perhaps representing normally subducted oceanic crust that was later uplifted and exposed (see Ernst, 1971, for a discussion of the Alps and similar eclogite occurrences in the Sanbagawa belt in Japan and the Franciscan belt in California). The metamorphic P-T conditions of the Alpine eclogites have been extensively studied, and it seems that eclogite persists in equilibrium to pressures of 10 kb at a temperature of ~500°C (Ernst, 1981; also see Brown and Bradshaw, 1979, for an analysis of the Franciscan eclogites). Therefore, it appears that the basalt→eclogite phase change will begin at a depth of 30-35 km in a subduction zone environment.

Transformational superplasticity

Under certain conditions, solids will deform to very large strains (>100%) when the applied stress is small, i.e. well below the elastic yield limit. This behavior is called superplasticity, and has been documented in the metallurgical literature (Johnson, 1970). Transformational plasticity is perhaps the most common type, and this occurs when a metal passes through a phase change. The apparent strength of the metal is substantially reduced as the crystal structure changes. To explain these observations, Greenwood and Johnson(1965) derived a theoretical result using a continuum formulation. The superplastic behavior is caused by the local volume change which results in high internal stresses, thereby placing the material in the plastic regime. Greenwood and Johnson showed that even if the volume-change induced stresses have no averaged preferred direction, the material responds to a small external stress by deforming plastically (note: this depends somewhat on the shape and behavior of the yield surface; Greenwood and Johnson assumed a Von Mises yield surface). The metallurgical experiments are conducted by cycling the temperature across a transformation temperature many times, thereby accumulating the strain. Superplasticity has been observed for phase changes with a volume change less than 1%.

The role of superplasticity in geophysics has received little attention. Sammis and Dein (1974) suggested that superplasticity should occur in silicates, and conducted an experiment that showed this

behavior in a silicate analog material (CsCl). Sammis and Dein were interested in applying superplasticity to the upper mantle discontinuities (at 400 and 650 km depth), as the intervening layers might be mechanically uncoupled across the assumed phase boundaries. One difficulty with this particular application is that a time continuous superplastic deformation requires the phase change to be occurring continuously in time, i.e. material has to be fluxing across the phase boundary. In our suggested application, there is new material constantly supplied to the depth interval of the phase change as the subducting slab provides a stream of oceanic crust.

We cannot yet prove that superplastic deformation will accompany the basalt→eclogite phase change, but it seems likely. One reservation is that while the phase changes in metals are relatively simple packing rearrangements, the basalt to eclogite phase change involves a multi-component system with new minerals appearing over a pressure range. This system is not exactly analogous to the metals. If one considers the superplastic process at the microscopic scale, a reasonable explanation for the plastic behavior is the enhanced mobility of dislocations during the phase change, as discussed by Poirier (1981). If this is the mechanism, then it should operate equally well in a more complicated multi-component system. In fact, Poirier (personal communication, 1981) thought it quite likely that superplastic deformation accompanies phase changes in silicate compounds.

Though basalt→eclogite superplasticity is likely, it does not immediately follow that the plates are completely seismically uncoupled

by this mechanism alone. If all of the relative plate motion is accommodated by shear deformation in the oceanic crust, the shear strain rate in the crust will be: $\dot{\varepsilon}=v/2h$, where v is the convergence velocity and h is the thickness of the superplastic zone. Assuming that the volume changes linearly in the phase change pressure interval: then the volume change rate is $\dot{\theta}=(.15)v/1$ with $\theta=(\Delta V/V)$ the volume change, v the convergence velocity, 1 the distance over which the phase change occurs, and (.15) is the total fractional volume change. Letting h= 6 km and 1= 30 km, we see that $\dot{\varepsilon}\sim16\dot{\theta}$. Hence, for complete uncoupling we require that the shearing strain rate be larger than the rate of volume change locally. This condition is unlike that assumed in the analysis of Greenwood and Johnson (1965). Though superplastic deformation will still occur, it is an open question whether the above strain rate will be obtained at low stress by superplastic deformation alone. Other processes which might accompany the phase change, such as the movement of water, may help in obtaining complete uncoupling.

A relevant observation is the distinct bend in the Benioff zone at a depth of 30 km in the Central Aleutians (see Figure 4.4, Engdahl, 1977). A bend has also been observed in the Shumagin region (Reyners and Coles, 1981), and there is a suggestion of a bend at ~30 km in Japan as well (see Yoshii,1979). This bend is resolved in the Aleutians and Japan due to the dense station distribution of the local networks, hence we cannot determine whether this is a general feature or not. The bend occurs at a depth where the basalt→eclogite transformation is expected to commence. Although it is uncertain what causes the bend, (e.g.
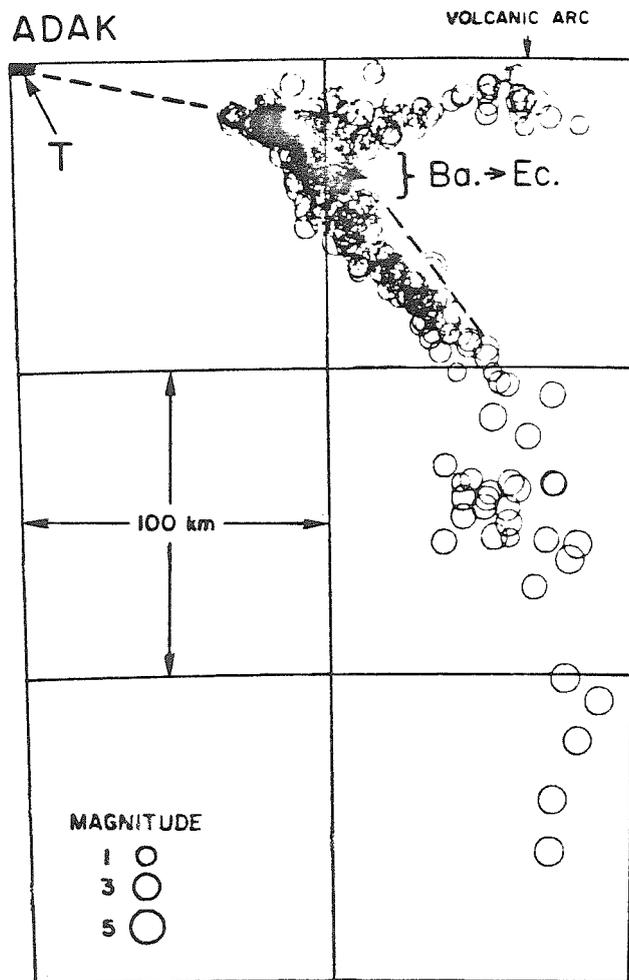
Figure 4.4. A cross-section of the seismicity in the Adak region of the Aleutians, after Engdahl (1977). Projecting the dips from the trench (T) and the deeper portion of the slab, a bend in the slab is required at a depth of ~30 km. The bend may occur as a distinct change in dip at the starting depth of the basalt to eclogite phase change.

increased density of slab, or perhaps an abrupt mechanical strength change), it does signify a distinct change in the subducting slab which corresponds to the limiting depth of seismic coupling.

In conclusion, it seems that there is a significant change in the plate coupling below a depth of 40 km, and a major phase change from basalt to eclogite affecting the 6 km thick oceanic crust can be expected to start at a depth of 30-35 km. We suggest that this phase change may cause the uncoupling. In particular, the oceanic crust should deform plastically while the phase change is occurring. Hence, superplastic shear strain throughout the oceanic crust could lead to at least partial uncoupling, but other processes associated with the basalt→eclogite transformation may be required to cause complete uncoupling. The superplastic deformation will occur only in the depth interval of the phase change. There is no direct information on the extent of the depth interval, and we have assumed that it is approximately 20 km. Thus, some other process may be involved to uncouple the plates below a depth of 60 km, although it could still be related to the basalt→eclogite phase change. For example, this phase change might expel water into the contact zone that decreases the coupling, or superplastic deformation causes the crust to be permanently weakened. We view the occurrence of the limiting depth of seismic coupling and the beginning of the basalt→eclogite phase change as highly suggestive that this phase change initiates the uncoupling of the plates.

4.4. Asperity Size and Plate Properties

Earthquake size is correlated to plate age and convergence velocity. We have hypothesized that earthquake size is related to asperity size. Therefore, we need to relate age and rate to asperity size. Establishing this relation is difficult as we do not know the specific physical nature of asperities. They could be either geometric irregularities of the plate surfaces, or varying strength along the narrow contact zone.

The simplest model is one in which the asperity size is directly determined by the preferred trajectory. This can occur in two ways: (i) the preferred trajectory determines the dip angle of the shallow contact plane, and the dip angle determines the width of the contact zone, (ii) if the shallow dip angle is fixed by the mechanics of bending, then preferred trajectory determines the horizontal compressive stress which locks or unlocks the surface irregularities.

If preferred trajectory is not directly responsible for asperity size, then there are many possibilities. We will discuss several possibilities here. Asperity size might be controlled by the intrinsic irregularities of the down-going plate, with an uneven surface causing small asperities. These irregularities may result from the topography before subduction. Even if the seamounts and ridges are eroded during subduction, the residual topography might still be important. Older oceanic lithosphere generally has a greater concentration of seamounts. Since seamounts represent small geometrical asperities, this mechanism

would explain the observed correlation.

Another source of geometric irregularities could be the horst and graben structure that appears in the oceanic plate at the trench (Hilde and Sharman, 1978). Displacements on these faults vary from 100 meters to over a kilometer, so that the retention of this relief during subduction might cause asperities. We contend that this relief is either more pronounced, or is better retained upon subduction, in oceanic lithosphere that is older and slowly subducting.

Asperities could also be caused by variations in the strength of the contact zone material. The subduction of sediments is then an important process; in particular the amount and type of sediments subducted are important. The plate properties might influence the amount of sediments subducted through differences in the horst and graben structure, as the volume of the sediments subducted may be controlled by the volume of the grabens (Hilde and Sharman, 1978). Alternatively, if the sediments are subducted as a thin veneer, the convergence rate along with the mechanical properties of the sediments may determine the thickness of the subducted sediment layer. Sediment supply, as well as type, can be strongly influenced by the distance from a continental landmass. Hence, proximity to a large landmass could affect the asperity distribution.

Regarding the direct effect of preferred trajectory, it is interesting to note the broadening of the shallow Benioff zone from the Aleutians to Alaska (Jacob et al., 1977). While this broadening can be explained by a change in the preferred trajectory from the Aleutians to

Alaska, it is also correlated with sediment supply. Also, the Alaskan rupture was unusually smooth and homogeneous. It is possible that the shallow dip is caused by sediment loading and the homogeneous rupture is due to an excessive thickness of subducted sediments. To further assess this possibility, we examine the setting of the 1960 Chile earthquake, the largest recorded earthquake. Indeed, it seems that the trenchline moves westward just at the northern edge of the 1960 rupture zone (Figure 4.5), and Kelleher et al.(1974) stated that the distance from the trench to the 70 km depth contour of the Benioff zone increases from Central Chile to Southern Chile. However, this change is not nearly as pronounced as the Aleutian-Alaska transition, and in fact the trench-volcano distance does not seem to change at all. Interpreting the bend in the trenchline is made more difficult by the fact that a fracture zone intersects the trench at this bend, with significantly younger lithosphere to the south. Thus, the rupture zone of the 1960 event and the apparent broadening of the shallow Benioff zone may reflect the younger lithosphere more than any other process. However, the trench sediment thickness is greater in the zone of the 1960 rupture (Kulm et al., 1977). This could be due to either increased sediment supply as the average annual rainfall on the coastal plain increases by a factor of four from Central Chile to Southern Chile, or that younger lithosphere subducts a smaller amount of sediments. The fact that both Alaska and Southern Chile have thicker sediments might lead us to conclude that excessive sediments somehow facilitate strong coupling. There are some counter-examples though: both the Caribbean and Sumatra
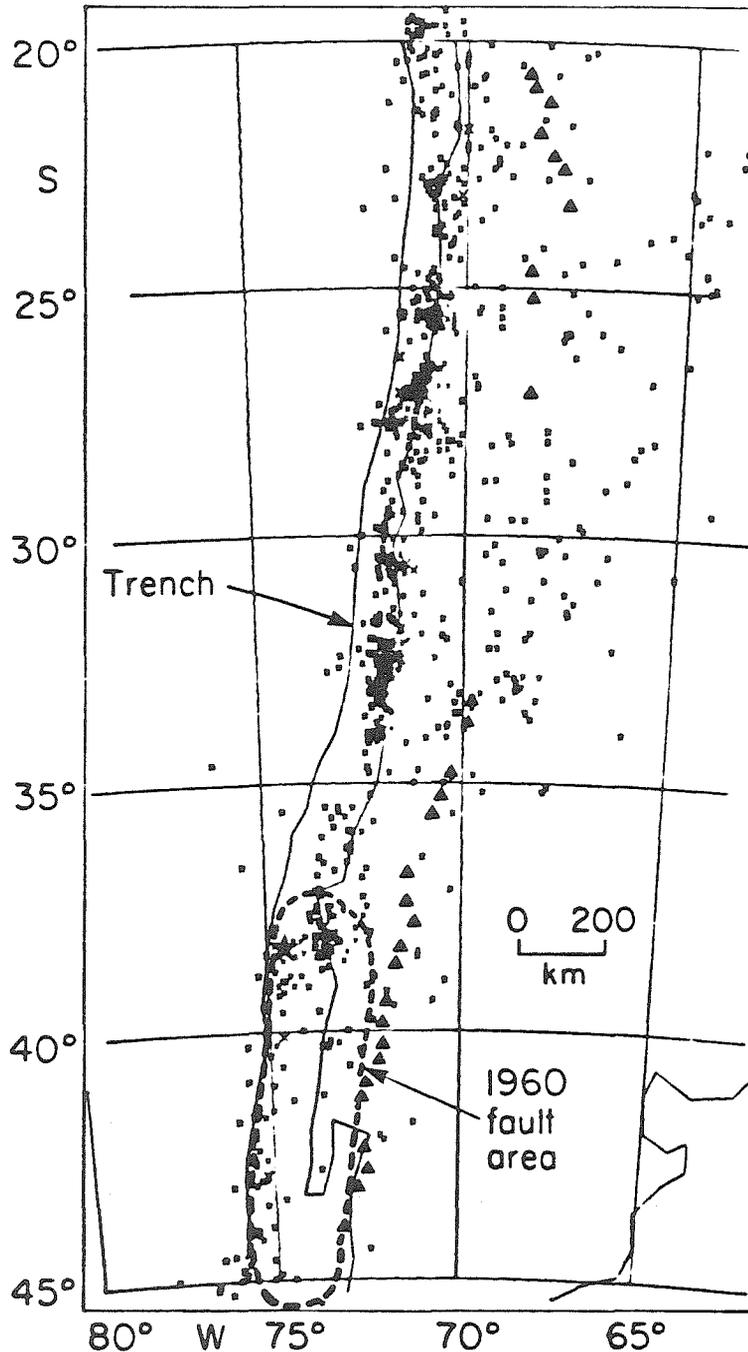
Figure 4.5. The seismicity (from 1963 to 1975), volcanos (triangles), and trench axis in the southern part of South America. The fault area and epicenter (star) of the May 22, 1960 event are shown. Note that the trench axis bends westward at the rupture zone of the 1960 event, and this coincides with an increased thickness of sediments. However, the trench-volcano distance does not seem to increase as in the Aleutians-Alaska transition. Also, a fracture zone intersects the trench around 35$\overset{\circ}{S}$.

zones have excessive sediments but no great earthquakes (in the southern Caribbean). One interpretation of the presence of excessive sediments at Alaska and Southern Chile is that plate age and rate determine the potential for earthquake size, while excessive sediments can enhance coupling by causing a smooth homogeneous contact surface.

In conclusion, the simplest relation between plate properties and earthquake size is preferred trajectory directly controlling asperity distribution through either the dip of the shallow contact zone, or the horizontal compressive stress influencing the contact between irregularities. Certainly this subject needs more work to isolate the important physical processes at subduction zones that cause earthquakes.

4.5. Summary

We have reviewed some of the correlations concerning the properties of subduction zones. A more general review of subduction zone comparisons can be found in Uyeda (1981). Our main interest here is in explaining the variation in characteristic earthquake size of subduction zones. There are three basic observations that we are attempting to explain: (i) the correlation of earthquake size to plate age and rate, (ii) the rupture processes of three great earthquakes which suggest that asperity size determines earthquake size, and (iii) all subduction zones appear to be uncoupled below a depth of 40 km. We present a simple model of seismic coupling that connects asperity size, earthquake size, and tectonic stress. However, relating asperity size to the plate age

and rate is rather difficult as we do not yet know the physical nature of asperities, nor what determines their size distribution. Given this level of uncertainty, we can only submit several proposals which hopefully can be tested. One question that must be addressed: are asperities dependent directly on preferred trajectory or are they influenced only through indirect means? The direct mechanisms that we have discussed are preferred trajectory controlling the shallow dip and locking or unlocking irregularities. The plate age and rate could also: determine oceanic plate topography causing many small asperities for older lithosphere, or affect the relief caused by bending at the trench. If subducted sediments are controlling asperities, plate properties can still be involved, though in a rather indirect manner.

Nearly all of the accumulated seismic slip at subduction zones is above a depth of 40 km. It seems that the expected occurrence of the basalt→eclogite phase change at about this depth is more than a coincidence. This phase change represents a major physical change in the subducting oceanic crust. We have suggested one mechanism, i.e. superplastic deformation throughout the oceanic crust, that might cause the uncoupling of the two plates. There could be other consequences of the phase change that would aid the uncoupling, such as the movement of water into the contact zone.

These ideas should be further tested to determine the overall viability of the seismic coupling model, and the model of asperities controlling earthquake size. The uncoupling aspect can be tested by deformation experiments and theoretical considerations of the possible

creep mechanisms, including superplasticity. The asperity model can be tested by more seismological studies of large subduction zone earthquakes. It is also important to understand what asperities actually are, and how their distribution is determined. This is a difficult problem, and hopefully the correlation of earthquake size to the plate properties offers some insight into the relevant processes.

References for Part I

Abe, K., 1975. Re-examination of the fault model for the Niigata earthquake of 1964. J. Phys. Earth, 23: 349-366.

Abe, K. and Kanamori, H., 1979. Temporal variation of the activity of intermediate and deep focus earthquakes. J. Geophys. Res., 84: 3589-3595.

Ahrens, T. J. and Schubert, G., 1975. Gabbro-eclogite reaction rate and its geophysical significance. Rev. Geophys. Space Phys., 13: 383-400.

Aki, K., 1966. Generation and propagation of G waves from the Niigata earthquake of June 16, 1964. Bull. Earthquake Res. Inst., Tokyo Univ., 44: 23-88.

Aki, K. and Richards, P. G., 1980. Quantitative Seismology. W. H. Freeman and Company, San Francisco, 932 pp.

Alewine, R. W., 1974. Application of Linear Inversion Theory Toward the Estimation of Seismic Source Parameters. Ph.D. Thesis, California Institute of Technology, 303 pp.

Algermissen, S. T., Rinehart, W. A., Sherburne, R. W., and Dillinger, W. H., 1969. Preshocks and aftershocks of the Prince William Sound earthquake of March 28, 1964. In: L. E. Leipold (Ed.), The Prince William Sound, Alaska Earthquake of 1964 and Aftershocks, Vol. II. U.S. Department of Commerce, Coast and Geodetic Survey, Washington D.C., 350 pp.

Anderson, D. L., 1979. The upper mantle transition region: eclogite? Geophys. Res. Lett., 6: 433-436.

Anderson, D., 1979b. The deep structure of continents. J. Geophys. Res., 84: 7555-7560.

Andrews, D. and Sleep, N., 1974. Numerical modelling of tectonic flow behind island arcs. Geophys. J. R. Astron. Soc., 38: 237-251.

Backus, G. and Gilbert, F., 1968. The resolving power of gross Earth data. Geophys. J. R. Astron. Soc., 16: 169-205.

Backus, G. and Gilbert, F., 1970. Uniqueness in the inversion of inaccurate gross Earth data. Phil. Trans. R. Soc. A, 266: 123-192.

Bearth, P., 1959. Uber eklogite, glaukophanschiefer und metamorphe Pillowlaven.Schweiz. Minerall. Petrog. Mitt., 39: 267-286.

Brown, E. H. and Bradshaw, J. Y., 1979. Phase relations of pyroxene and amphibole in greenstone, blueschist, and eclogite of the Franciscan complex, California. Contrib. Mineral. Petrol., 71: 67-83.

Cohen, L. H., Ito, K. and Kennedy, G. C., 1967. Melting and phase relations in an anhydrous basalt to 40 kilobars. Am. J. Sci., 265: 475-518.

Doornbos, D. J. and Mondt, J. C., 1979. P and S waves diffracted around the core and the velocity structure at the base of the mantle. Geophys. J. R. Astron. Soc., 57: 381-395.

Eguchi, T., Uyeda, S. and Maki, T., 1979. Seismotectonics and tectonic history of the Andaman Sea. Tectonophysics, 57: 35-51.

Elasser, W., 1971. Sea-floor spreading as convection. J. Geophys.

Res., 76: 1101-1112.

Engdahl, E. R., 1977. Seismicity and plate subduction in the Central Aleutians. In: M.Talwani and W.C. Pitman III (Eds.), Island Arcs, Deep Sea Trenches, and Back-Arc Basins. American Geophysical Union, Washington, D.C., 480 pp.

Ernst, W. G., 1971. Metamorphic zonations on presumably subducted lithospheric plates from Japan, California, and the Alps. Contrib. Mineral. Petrol., 34: 43-59.

Ernst, W. G., 1981. Petrogenesis of eclogites and peridotites from the Western and Ligurian Alps. Am. Mineral., 66: 443-472.

Green, D. H. and Ringwood, A. E., 1967. An experimental investigation of the gabbro to eclogite transformation and its petrological applications. Geochim. Cosmochim. Acta, 31: 767-833.

Greenwood, G. W. and Johnson, R. H., 1965. The deformation of metals under small stesses during phase transformations. Proc. Roy. Soc., A283: 403-422.

Gutenberg, B. and Richter, C. F., 1954. Seismicity of the Earth, 2nd ed., Princeton University Press, Princeton, N.J., 310 pp.

Hager, B. H. and O'Connell, R. J., 1978. Subduction zone dip angles and flow driven by plate motion. Tectonophysics, 50: 111-133.

Hastie, L. M. and Savage, J. C., 1970. A dislocation model for the Alaska earthquake. Bull. Seism. Soc. Am., 60: 1389-1392.

Herrin, E. (Chairman), 1968. 1968 Seismological Tables for P Phases. Bull. Seism. Soc. Am., 58: 1193-1241.

Hilde, T. W. C. and Sharman, G. F., 1978. Fault structure of the descending plate and its influence on the subduction process. (Abstract) EOS, 59: 1182.

Hirasawa, T., 1965. Source mechanism of the Niigata Earthquake of June 16, 1964, as derived from body waves. J. Phys. Earth, 13: 35-66.

House, L. and Boatwright J., 1980. Investigation of two high stress drop earthquakes in the Shumagin seismic gap, Alaska. J. Geophys. Res., 85: 7151-7165.

Hsui, A. T. and Toksoz, M. N., 1979. The evolution of thermal structures beneath a subduction zone. Tectonophysics, 60: 43-60.

Isacks, B., Oliver, J. and Sykes, L.R., 1968. Seismology and the new global tectonics. J. Geophys. Res., 73: 5855-5899.

Isacks, B., Sykes, L. and Oliver, J., 1969. Focal mechanisms of deep and shallow earthquakes in the Tonga-Kermedec region and the tectonics of island arcs. Bull. Geol. Soc. Amer., 80: 1443-1470.

Jackson, D. D., 1972. Interpretation of inaccurate, insufficient, and inconsistent data. Geophys. J. R. Astron. Soc., 28: 97-109.

Jacob, K. H, Nakamura, K. and Davies, J. N., 1977. Trench-volcano gap along the Alaska-Aleutian arc: facts and speculations on the role of terrigeneous sediments. In: M. Talwani and W.C. Pitman III (Eds.), Island Arcs, Deep Sea Trenches, and Back-Arc Basins. American Geophysical Union, Washington, D.C., 480 pp.

Johnson, R. H., 1970. Superplasticity. Met. Rev., 146: 115-134.

Kanamori, H., 1970a. Synthesis of long-period surface waves and its
   application to earthquake source studies -- Kurile Islands
   earthquake of October 13, 1963. J. Geophys. Res., 75:
   5011-5027.

Kanamori, H., 1970b. The Alaska earthquake of 1964: Radiation of
   long-period surface waves and source mechanism. J. Geophys.
   Res., 75: 5029-5040.

Kanamori, H., 1971. Great earthquakes at island arcs and the
   lithosphere. Tectonophysics, 12: 187-198.

Kanamori, H., 1977a. The energy release in great earthquakes. J.
   Geophys. Res.,82: 2981-2987.

Kanamori, H., 1977b. Seismic and aseismic slip along subduction zones
   and their tectonic implications. In: M. Talwani and W.C.
   Pitman III (Eds.), Island Arcs, Deep Sea Trenches, and Back-Arc
   Basins. American Geophysical Union, Washington, D.C., 480 pp.

Kanamori, H., 1978. Quantification of earthquakes. Nature, 271:
   411-414.

Kanamori, H., 1981. The nature of seismicity patterns before large
   earthquakes. In: D.W. Simpson and P.G. Richards (Eds.),
   Earthquake Prediction, an International Review. American
   Geophysical Union, Washington, D.C., 680 pp.

Kanamori, H. and Anderson, D. L., 1975. Theoretical basis of some
   empirical relations in seismology. Bull. Seism. Soc. Am., 65:
   1073-1095.

Kanamori, H. and Cipar, J. J., 1974. Focal process of the great

Chilean earthquake May 22, 1960. Phys. Earth Planet. Int., 9: 128-136.

Kanamori, H. and Given, J. W., 1981. Use of long-period surface waves for rapid determination of earthquake-source parameters. Phys. Earth Planet. Int., 27: 8-31.

Kanamori, H. and McNally, K. C., 1981. Variable rupture mode of the subduction zone along the Ecuador-Colombia coast. preprint.

Kanamori, H. and Stewart, G. S., 1976. Mode of the strain release along the Gibbs fracture zone, mid-Atlantic ridge. Phys. Earth Planet. Int., 11: 312-332.

Karig, D., 1971. Origin and development of marginal basins in the Western Pacific. J. Geophys. Res., 76: 2542-2561.

Kelleher, J. and McCann, W., 1976. Buoyant zones, great earthquakes, and unstable boundaries of subduction. J. Geophys. Res., 81: 4885-4896.

Kelleher, J., Savino, J., Rowlett, H. and McCann, W., 1974. Why and where great thrust earthquakes occur along island arcs. J. Geophys. Res., 79: 4889-4899.

Kulm, L. D., Schweller, W. J., and Masias, A., 1977. A preliminary analysis of the subduction processes along the Andian continental margin, $6°$ to $45°$S. In: M. Talwani and W.C. Pitman III (Eds.), Island Arcs, Deep Sea Trenches, and Back-Arc Basins. American Geophysical Union, Washington, D.C., 480 pp.

Lanczos, C., 1961. Linear Differential Operators, D. Van Nostrand Co., London.

Langston, C. A. and Helmberger, D. V., 1975. A procedure for modeling shallow dislocation sources. Geophys. J. R. Astron. Soc., 42: 117-130.

Lay, T. and Kanamori, H., 1980. Earthquake doublets in the Solomon Islands. Phys. Earth. Planet. Int., 21: 283-304.

Lay, T. and Kanamori, H., 1981. An asperity model of great earthquake sequences. In: D.W. Simpson and P.G. Richards (Eds.), Earthquake Prediction, an International Review. American Geophysical Union, Washington, D.C., 680 pp.

Lister, C. R. B., 1977. Qualitative models of spreading-center processes, including hydrothermal penetration. In: S. Uyeda (Ed.), Subduction Zones, Mid-Ocean Ridges, Oceanic Trenches and Geodynamics. Tectonophysics, 37: 203-218.

Luyendyk, B. P., 1970. Dips of downgoing lithospheric plates beneath island arcs. Geol. Soc. Am. Bull., 81: 3411-3416.

MacKenzie, D., 1969. Speculations on the consequences and causes of plate motions. Geophys. J. R. Astron. Soc., 18: 1-32.

Minster, J., Jordan, T., Molnar, P. and Haines, E., 1974. Numerical modelling of instantaneous plate tectonics. Geophys. J. R. Astron. Soc., 36: 541-576.

Mikumo, T. and Miyatake, T., 1978. Dynamical rupture process on a three-dimensional fault with non-uniform frictions and near-field seismic waves. Geophys. J. R. Astron. Soc., 54: 417-438.

Miyashita, K. and Matsuura, Y., 1976. Dislocation distribution on the fault associated with the Alaska earthquake of 1964 (in Japanese).

Paper presented at Annual Spring Meeting, Seism. Soc. of Japan, Tokyo.

Molnar, P. and Atwater, T., 1978. Inter-arc spreading and cordilleran tectonics as alternates related to the age of subducted oceanic lithosphere. Earth Planet. Sci. Lett., 41: 330-340.

Molnar, P. Freedman, D. and Shih, J., 1979. Lengths of intermediate and deep seismic zones and temperatures in downgoing slabs of lithosphere. Geophys. J.R. Astron. Soc., 56: 41-54.

Mula, A. H. G., 1981. Amplitudes of diffracted long-period P and S waves and the velocities and Q structure at the base of the mantle. J. Geophys. Res., 86: 4999-5011.

Plafker, G., 1972. Alaskan earthquake of 1964 and Chilean earthquake of 1960: Implications for arc tectonics. J. Geophys. Res., 77: 901-925.

Poirier, J. P., 1981. On the kinetics of olivine-spinel transition. Phys. Earth Planet. Int., 26: 179-187.

Reyners, M. and Coles, K. S., 1981. Fine structure of the dipping seismic zone and subduction mechanics in the Shumagin Islands, Alaska. J. Geophys. Res. (in press).

Ringwood, A. E. and Green, D. H., 1966. An experimental investigation of the gabbro-eclogite transformation and some geophysical implications. Tectonophysics, 3: 383-427.

Rudnicki, J.W. and Kanamori, H., 1981. Effects of fault interaction on moment, stress drop, and strain energy release. J. Geophys. Res., 86: 1785-1793.

Sammis, C. G. and Dein, J. L., 1974. On the possibility of transformational superplasticity in the Earth's mantle. J. Geophys. Res., 79: 2961-2965.

Seno, T., 1977. The instantaneous rotation vector of the Philippine Sea Plate relative to the Eurasian plate, Tectonophysics, 42: 209-226.

Sleep, N. and Toksoz, M., 1971. Evolution of marginal basins. Nature, 33: 548-550.

Stewart, G. S., Chael, E. P., and McNally, K. C., 1981. The November 29, 1978, Oaxaca, Mexico, earthquake: A large simple event. J. Geophys. Res., 86: 5053-5060.

Stewart, G. S. and Cohn, S. N., 1979. The 1976 August 16, Mindanao, Philippine earthquake (M_s = 7.8) -- Evidence for a subduction zone south of Mindanao. Geophys. J. R. Astron. Soc., 57: 51-65.

Sykes, L. R. and Quittmeyer, R. C., 1981. Repeat times of great earthquakes along simple plate boundaries. In: D.W. Simpson and P.G. Richards (Eds.), Earthquake Prediction, an International Review. American Geophysical Union, Washington, D.C., 680 pp.

Toksöz, M. and Hsui, A., 1978. Numerical studies of back-arc convection and the formation of marginal basins. Tectonophysics, 50: 177-196.

Tovish, A., Schubert, G., and Luyendyk, B. P., 1978. Mantle flow pressure and the angle of subduction: non-Newtonian corner flows. J. Geophys. Res., 83: 5892-5898.

Uyeda, S., 1977. Some basic problems in the Trench-Arc-Back arc system.

In M. Talwani and W. Pitman III (Eds.), Island Arcs, Deep Sea
Trenches and Back-arc Basins, American Geophysics Union,
Washington D.C., 480 pp.

Uyeda, S., 1981. Subduction zones (Introduction to comparative
subductology). Submitted to Proc. 26th Int. Geol. Congress.

Uyeda, S. and Kanamori, H., 1979. Back-arc opening and the mode of
subduction. J. Geophys. Res., 84: 1049-1061.

Vlaar, N. J. and Wortel, M. J. R., 1976. Lithospheric aging,
instability, and subduction. Tectonophysics, 32: 331-351.

Wiggins, R. A., 1972. The general linear inverse problem:
Implications of surface waves and free oscillations for Earth
structure. Rev. Geophys. Space Phys., 10: 251-285.

Wortel, M. and Vlaar, N., 1978. Age-dependent subduction of oceanic
lithosphere beneath Western South America. Phys. Earth Planet.
Int., 17: 201-208.

Wu, F., 1972. The Philippine Sea Plate: A "sinking towel"?,
Tectonophysics, 14: 81-86.

Wu, F. T. and Kanamori, H., 1973. Source mechanism of February 4,
1965 Rat Island Earthquake. J. Geophys. Res., 78: 6082-6092.

Wyllie, P. J., 1971. The Dynamic Earth. J. Wiley & Sons, New York,
416 pp.

Wyss, M. and Brune, J. N., 1967. The Alaska earthquake of 28 March
1964: A complex multiple rupture. Bull. Seism. Soc. Am., 57:
1017-1023.

Yoder, H. S. and Tilley, C. E., 1962. Origin of basalt magmas: an

experimental study of natural and synthetic rock systems. J. Petrol., 3: 342-532.

Yoshii, T., 1979. A detailed cross-section of the deep seismic zone beneath northeastern Honshu, Japan. Tectonophysics, 55: 349-360.

Appendix A for Part I

The purpose of this appendix is to investigate further the characterization of coupling strength. In the main text, we take the size of the largest earthquake as representative of the coupling strength. Perhaps it is better to use the cumulative seismic moment. That is, a more appropriate physical measure of seismic coupling would be the cumulative moment per unit time per unit subduction zone trench length. While it is easy to correct for subduction zone trench length, the moment release per unit time is difficult to estimate, as a regular recurrence interval and a complete sequence of events are required. The uncertainties in the recurrence intervals discourage any attempt to systematically apply corrections. If we were to normalize the seismic moment to a recurrence interval of 100 years, the zones that would be affected the most are those in the lower left-hand corner of Figure 2.2. It is quite possible that the recurrence interval for some of these regions is ~100 y or longer. Given these uncertainties, we are not able to apply corrections for the recurrence interval.

We use a cumulative moment, M_W', defined as

$$M_W' = (1/1.5) \log \Sigma \ 10^{1.5M_W},$$

summed over the known events in one sequence and corrected to subduction zone trench length = 1000 km. The only zones in which the increase over the single event M_W is 0.3 units or larger are: Peru, Central America, and Kuriles. These changes partially alleviate the low M_W values of Peru and Central America. The subduction zones and modified M_W values are plotted in Figure A1. As can be seen, the modified M_W values tend

to be more consistent with a plane fit, and the correlation coefficient, $r_{S,AR}$ is improved slightly to 0.844. In conclusion, Peru and Central America are more consistent with the general trend when using $M_{W'}$, though they are still low. These changes do not affect the conclusions in the main text.
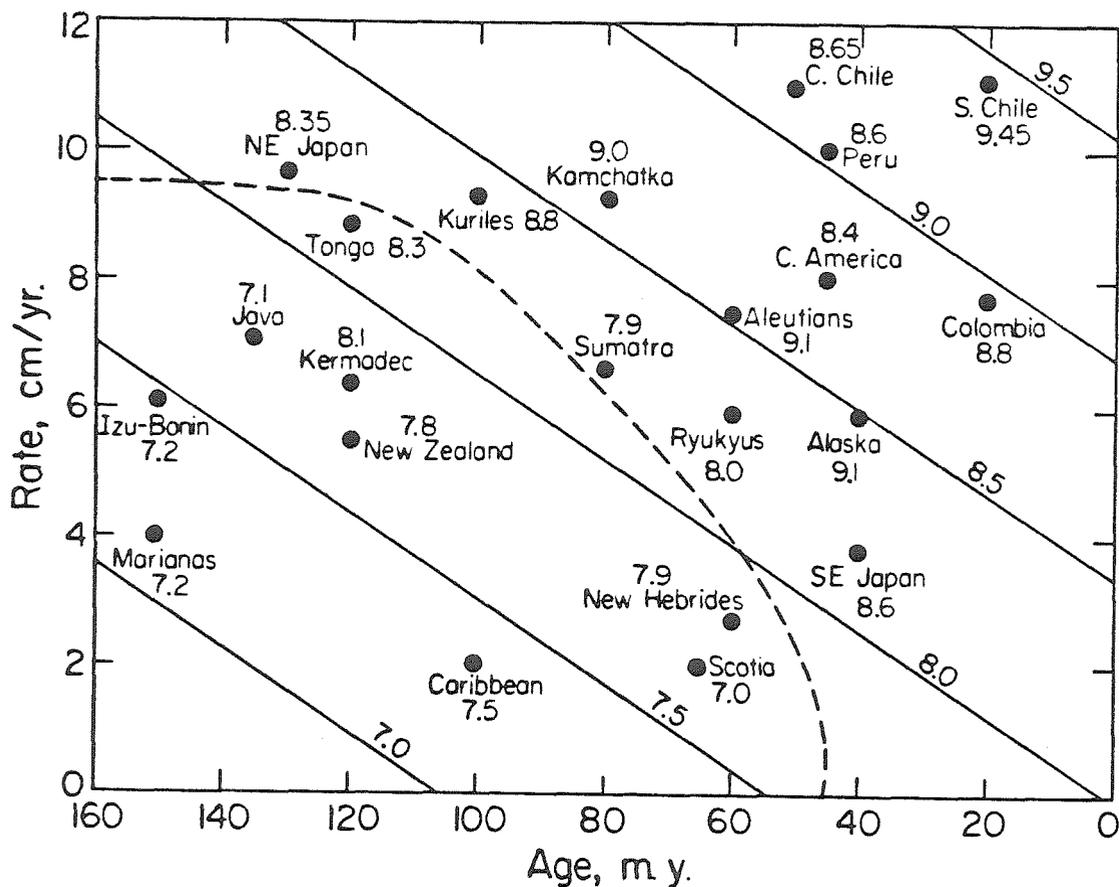
Figure A1. As Fig. 2.2, except that the cumulative moment magnitude, M_W', is used instead of the single event M_W. Also, the southeast Japan subduction zone has been added as the three variables: age, rate, and seismicity, can be determined. With seismicity represented by M_W', the resultant regression plane fits the data slightly better than in Figure 2.2.

Part II

Chapter 1

The Structure of the Lowermost Mantle Determined

by Short Period P-wave Amplitudes

1.1. Introduction

The lowermost mantle

The lowermost mantle has been a seismologically interesting region since the early Earth models of Jeffreys-Bullen and Gutenberg, which indicated anomalous velocity gradients for both P and S waves in the lowermost ~ 200 km of the mantle (hereafter referred to as the D" layer). Many different velocity models have been proposed for the lowermost mantle and there have also been various suggestions of lateral variations within the D" layer. The models can be divided into two categories: (i) smooth models, which allow an abrupt change in the velocity gradient at the top of D", but then have a nearly constant gradient extending down to the CMB (core-mantle boundary); (ii) rough models, which in addition to the change in gradient at the top of D", have a more complicated velocity distribution within D", e.g., a low velocity layer, a high velocity layer, or some combination. In addition, a complete description of D" should include the anelastic structure as there have been suggestions of a low-Q zone in the lower mantle.

Recent considerations of mantle dynamics have indicated that if heat from the core contributes significantly to the convective heat flux of the mantle, then a thermal boundary layer must exist at the base of the mantle so that the heat can diffuse into mantle material across the

CMB (Turcotte and Oxburgh, 1967; Jones, 1977; Sharpe and Peltier, 1979; Elasser, Olson, and Marsh, 1979; Jeanloz and Richter, 1979). Thus it is important to determine the existence or absence of a thermal boundary layer at the CMB. The temperature increase and thickness of the thermal boundary will vary regionally, and mapping out lateral variations in the elastic properites could detail the geometry of the boundary layer. However, the regionally-averaged properties of D'' should be determined first since the dominant effect of a thermal boundary layer will be to decrease the seismic velocities of the averaged structure due to the superadiabatic temperature gradient. As a note of caution, even if the averaged seismic velocities in D'' are determined to be anomalously low, consistent with a superadiabatic temperature gradient, one cannot simply conclude that a thermal boundary layer exists at the CMB. A compositional change would be a viable alternative and there have in fact been suggestions of a compositional change in D'': Bullen's (1950) proposal was based on the (k,p) hypothesis and more recently, a compositional difference has been proposed based on considerations of a particular core model (Anderson, 1975; Ruff and Anderson, 1980).

D" velocity models

The velocity structure can be determined throughout most of the mantle by inverting the travel times and $dT/d\Delta$ (inverse of apparent phase velocity). This technique is not adequate for the determination of velocity in D". The $dT/d\Delta$ of both P and S waves becomes nearly constant beyond the distance of $\Delta = 90$ deg and it is primarily for this reason that travel time and $dT/d\Delta$ studies invariably propose a smooth-type velocity structure in D". Though there is no general agreement on the particular value of the velocity gradient, the different studies agree on the constraint imposed on the velocity in D": the P-wave velocity does not exceed a constant value. The nature of this constraint can be seen in Figure 1.1. Earth models based on the inversion of normal modes also portray D" with a smooth structure, though this may not be additional evidence as their starting models are smooth and the normal modes used are probably sensitive only to the averaged velocity of D". The review article by Cleary (1974) discusses these issues in more detail.

Besides travel times and $dT/d\Delta$, other data that can resolve the structure of D" are the amplitudes of P and S waves bottoming in D" and the phases PcP, PKP, and SKS. Some studies that use the amplitudes of various phases that pass through D" call upon a rough structure to explain anomalies within their data. Two particular examples are: Mitchell and Helmberger (1973), who needed a thin high-velocity layer based on the amplitude ratios of S-ScS, and Bolt (1972), who suggested
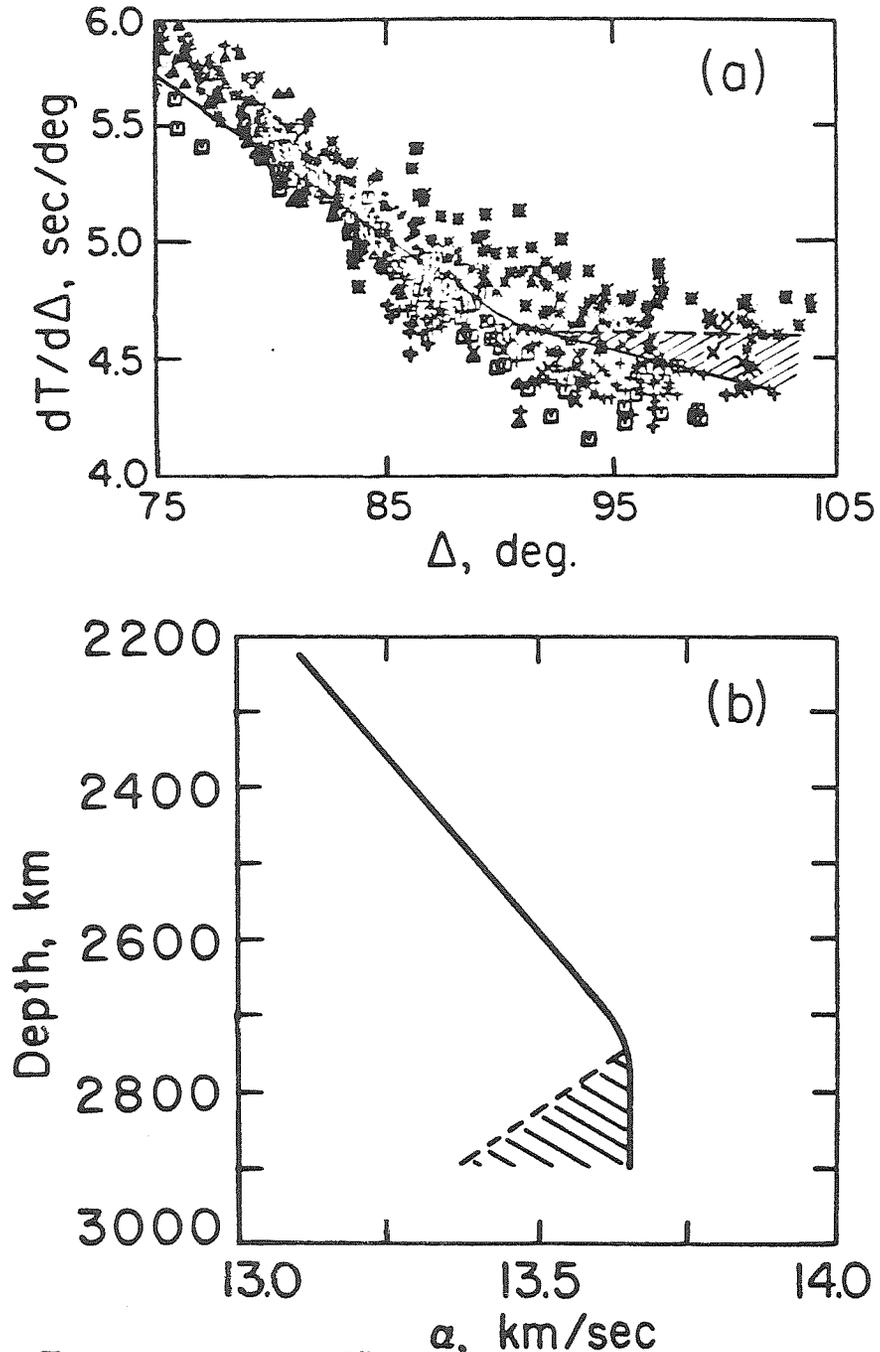
Figure 1.1. The uncertainty in D″ velocity structure due to a nearly constant dT/dΔ beyond Δ = 90 deg. In (a), dT/dΔ data relevant to the lower mantle are shown. The data were collected by C. Powell using the Caltech array, and the different symbols indicate different azimuths. The increased scatter beyond Δ ~ 90 deg. is apparent. The uncertainty in the velocity structure is indicated in (b), where the dashed and solid lines correspond to those in (a). The velocity in D″ could be anywhere within the shaded region.

that his short period P-wave amplitudes required a more complicated structure in D". The proposed rough models bear little resemblance to each other, hence no representative rough model has yet emerged.

To reconcile these two types of velocity models, smooth and rough, one might invoke a laterally varying D". This may not be necessary if we recognize that travel time, $dT/d\Delta$, normal mode, and long period P wave amplitude studies may not be capable of resolving the layered structure of D". A rough model might appear to be smooth to the above-mentioned methods, but produce short period amplitudes inconsistent with a smooth model. Thus, a careful study of short period amplitudes could resolve the internal structure of D", including the possibility of a low-Q zone (Teng, 1968; Anderson and Hart, 1978). A study of one particular region would avoid the additional problem of regional variations, leaving the question open until we can understand the structure in one locality. We have used short period P-wave amplitudes to study a relatively small region of the lowermost mantle.

Diffracted P waves

Theoretical studies (see Chapman and Phinney, 1972 for a review), have shown that the amplitude decay into the shadow zone is directly related to the velocity gradient in D", with a larger velocity gradient (increasing velocity with depth) causing a faster decay of amplitudes. As the period of the diffracting wave increases, the wave will average the velocity structure over an increasing thickness (progressing upward

from the CMB). There have been studies using the amplitude decay of long period (T = 8 to 64 seconds) P waves (Phinney and Alexander, 1969; Alexander and Phinney, 1966; Phinney and Cathles, 1969; Chapman and Phinney, 1972), but their results are inconclusive. The most complete effort using long period data is by Doornbos and Mondt (1979a,b), concluding that both P and S wave velocity gradients become negative with depth. Their testing of different models also demonstrates the limited resolution of the long period data, e.g., their long period data do not allow for any conclusion concerning Q.

The amplitudes of short period p waves are more sensitive to both the Q and velocity structure. There are only a few studies that have attempted to model short period amplitudes (Gutenberg, 1960; Teng, 1966; Phinney and Cathles, 1969; Bolt, 1972) and the modeling has been limited to tests of compatibility. The primary reason for lack of quantitative short period studies has been the absence of a high quality short period data set, as synthetic seismograms of diffracted waves can now be calculated. Previous compilations of short period amplitudes from D", such as Sacks (1966), Carpenter, Marshall and Douglas (1967) and Cleary (1967) have had a large uncertainty in the amplitude decay which has discouraged any quantitative comparisons. By a careful analysis of P wave amplitudes from underground nuclear explosions, this paper presents the best D" short period amplitude data set yet compiled, upon which useful quantitative comparisons can be based.

1.2 Short Period Amplitudes

While short period amplitudes are sensitive to the lowermost mantle structure, they are also sensitive to source and receiver effects. Source effects, such as radiation pattern and directivity, and receiver effects, such as differences in apparent attenuation, combine to greatly increase the scatter in measured amplitudes. This scatter may, in particular geometries, introduce systematic effects to a profile of amplitude versus distance. Therefore, to obtain a better quality data set, we need to minimize both the source and receiver effects. The best way to minimize source effects is to use underground nuclear explosions as the sources instead of earthquakes. The radiation pattern is then nearly spherically symmetric, there is no directivity, and there are usually many events confined to a small region which allows the measurements to be repeated. To correct for receiver effects, it is necessary to employ a suite of stations comparing the relative (i.e., to each other) station amplitudes from different source regions. After normalizing out the absolute amplitudes, the difference between the averaged amplitude value at each station and the mean for the suite of stations is designated as the receiver effect. A fortuitous arrangement of Russian test sites relative to the North America WWSSN stations (see Figure 1.2) allows for the determination of receiver effects using the northern test sites (Novaya Zemlya). These corrections can then be applied to the amplitudes from the southern test sites, thereby removing the common receiver effects.
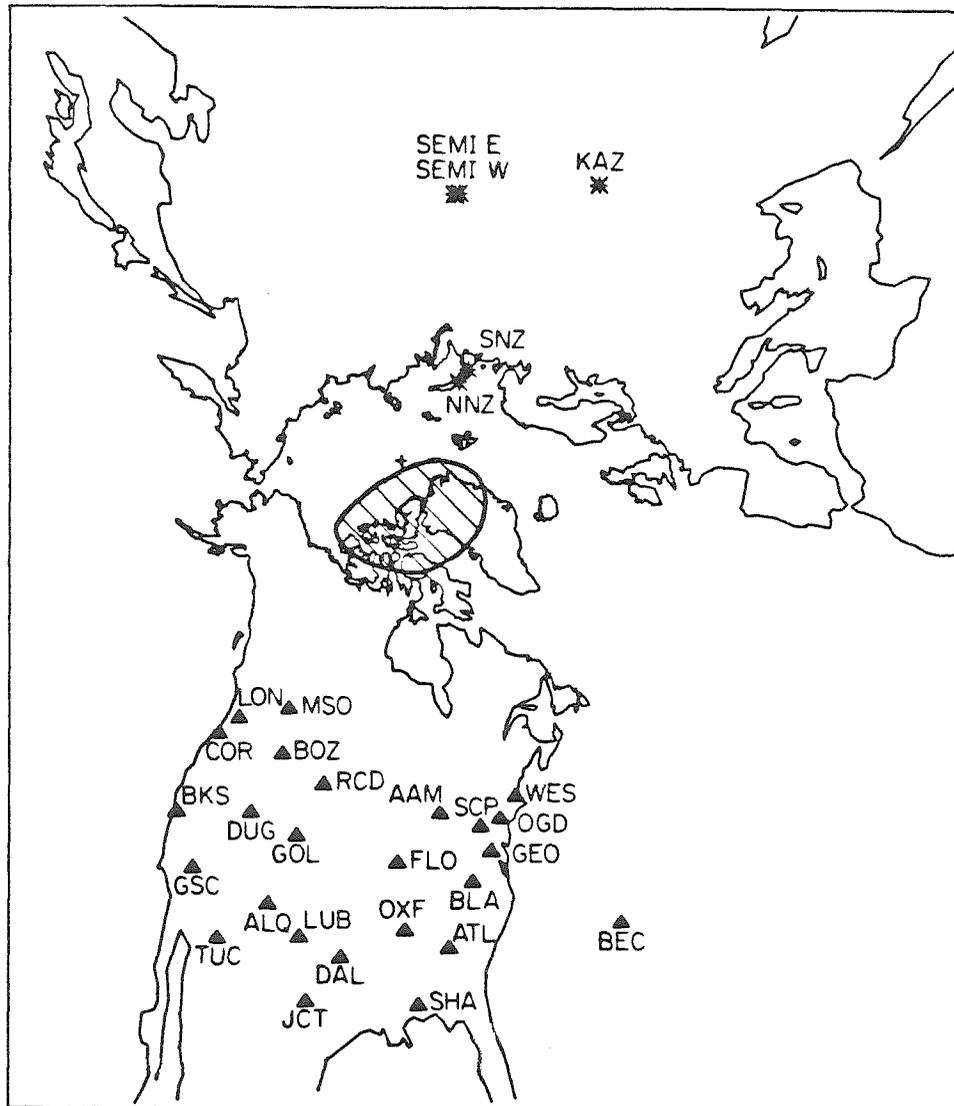
Figure 1.2. Gnomic projection of test sites and WWSSN stations (great circle paths are straight lines). Paths from Kazakh and Semipalatinsk to WWSSN stations sample the lowermost mantle beneath the shaded region.

TABLE 1.1
Distances between the WWSSN stations and the Semipalatinsk
and Kazakh test sites (Δ, deg.).

Station	SEMI	KAZ
AAM	86.9	86.5
ALQ	95.4	99.8
ATL	95.7	94.5
BEC	91.9	86.2
BKS	90.6	98.6
BLA	91.3	89.7
BOZ	84.4	90.1
COR	83.9	91.9
DAL	97.5	99.4
DUG	89.6	95.6
FLO	91.1	91.9
GEO	89.1	86.8
GOL	90.7	94.9
GSC	94.0	100.9
JCT	99.9	102.5
LON	82.0	89.7
LUB	96.8	100.1
MSO	83.0	89.2
OGD	86.5	84.0
OXF	95.3	95.5
RCD	86.3	90.9
SCP	87.4	85.6
SHA	98.9	98.4
TUC	97.6	103.1
WES	84.5	81.5

A more complete discussion of the basic data set and the procedure used to normalize the amplitudes is contained in Butler and Ruff (1980) and Ruff and Butler (1980). We should note that our convention of amplitude measurement is to use the first peak to trough amplitude. Our experience with nuclear explosion waveforms has shown this measurement to be least susceptible to both record noise and waveform distortion due to distinct near-receiver arrivals. The relation between this measure of amplitude and the overall waveform is discussed in the next section.

Amplitudes and waveforms

This paper is primarily concerned with amplitude as a function of distance. It is necessary though, to consider the waveforms to insure that the AB (first peak to trough) amplitude measure represents the overall amplitude of the waveform. There are three lines of evidence that demonstrate that the AB measurement is a stable quantity and a good characterization of the wave amplitude at a period of T=1 sec. The first supporting evidence is entirely observational: WWSSN short period recordings of nuclear events have a characteristic waveform that maintains the same peak to peak amplitude ratios despite a large variation in the absolute amplitude. This similarity can be seen in Figure 1.3. Given this feature, either the first peak to trough (AB) amplitude or the first trough to second peak (BC) amplitude, or their combination, would be an adequate measure of the whole waveform amplitude. We have used the AB amplitude to minimize any source and/or
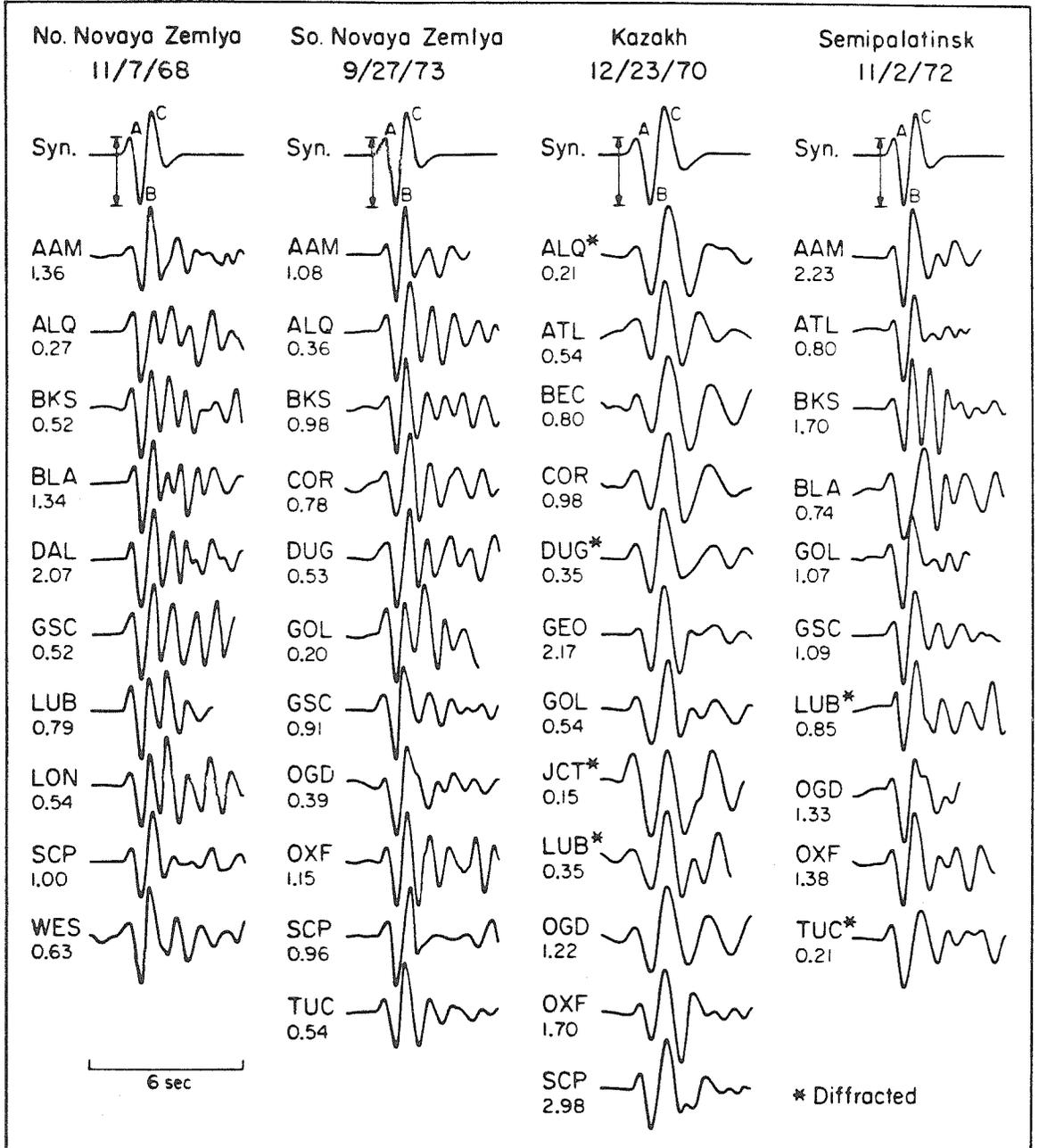
Figure 1.3. WWSSN short period seismograms of P waves for events at four different test sites. The first ~ 2 seconds of the seismograms are quite coherent, and this characterisitic waveform is easily reproduced by synthetic seismograms. The AB amplitude measure is indicated. The numbers give the relative amplitudes of the seismograms.

receiver secondary arrivals that would tend to affect the BC amplitude more than the AB value. The second line of evidence, which vouches for amplitude stability, is that upon processing the amplitudes there is a consistent pattern of receiver values (e.g., Figure 1.4). Though only an argument of self-consistency, this verification is important as otherwise there would be no well-defined receiver correction. The third line of evidence is that the observed characteristic waveform is easily matched by synthetic seismograms (e.g., Burdick and Helmberger, 1979). At distances between $\Delta = 30$ and $\Delta = 85$ the waveforms can be synthesized simply by convolving the instrument response with an attenuation operator and the source time function (including pP). Thus, the observed waveform is mostly explained by the source and attenuation.

Presumably the earth response causes the amplitude decay into the shadow and we would expect the waveforms to be modified also. The change in overall amplitude is more pronounced than the change in waveshape, due in part to the bandwidth of the WWSSN instrument (see Figure 1.3). If we wish to test earth models, we need to synthesize the waveforms and then measure the AB amplitude of the synthetic seismograms. A spectral method is not adequate as the waveshape does slowly change in the diffracted region, and the dominant period of the AB measure gradually changes to $T{\sim}1.4$ sec at $\Delta{\sim}104$. Therefore, to properly model the amplitude profile, reliable synthetics must be constructed. Although we only use the AB value of the synthetics, the waveforms are checked for their agreement with the observed waveforms. Summarizing this section, the observed waveforms and AB amplitudes
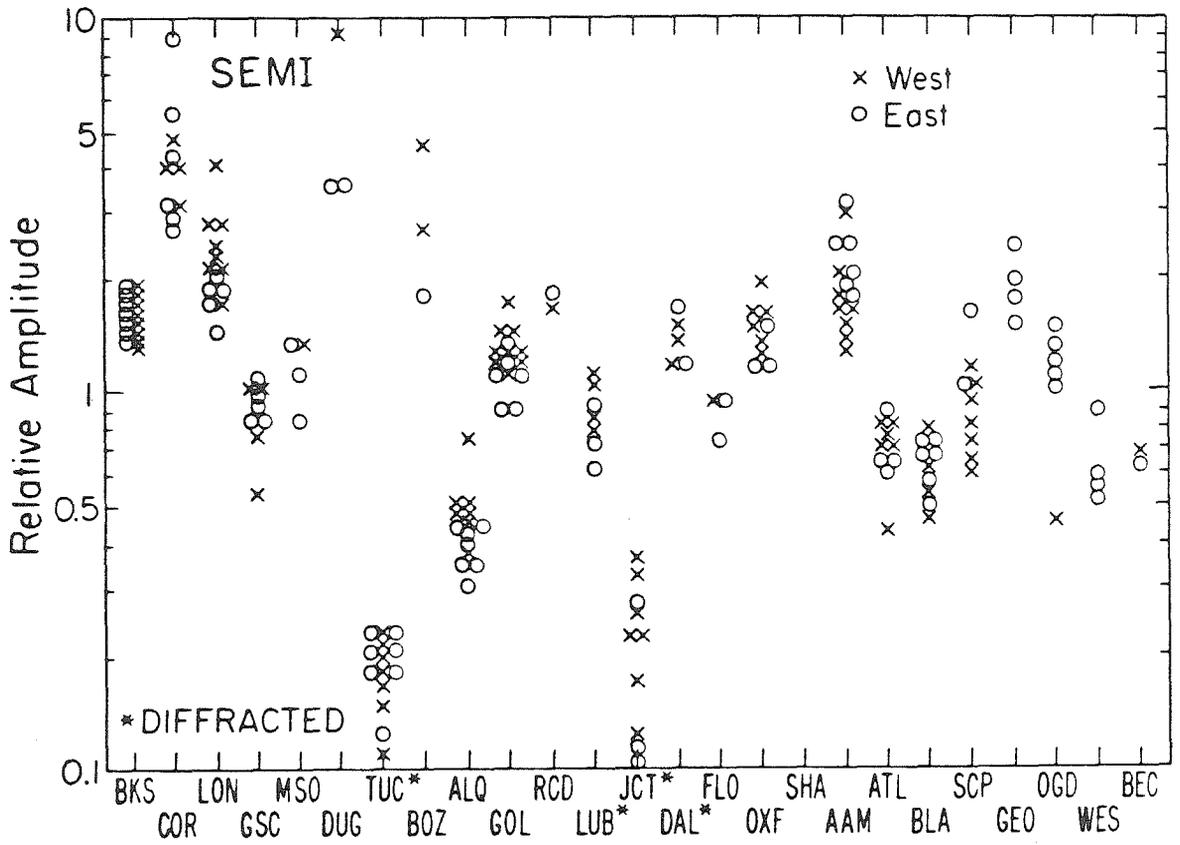
Figure 1.4. The relative AB amplitudes recorded at the WWSSN stations for events at the Semipalatinsk sites. The absolute magnitude of each event is adjusted to minimize scatter about the mean station amplitudes. (All of the events used at each test site are listed in Butler and Ruff, 1980).
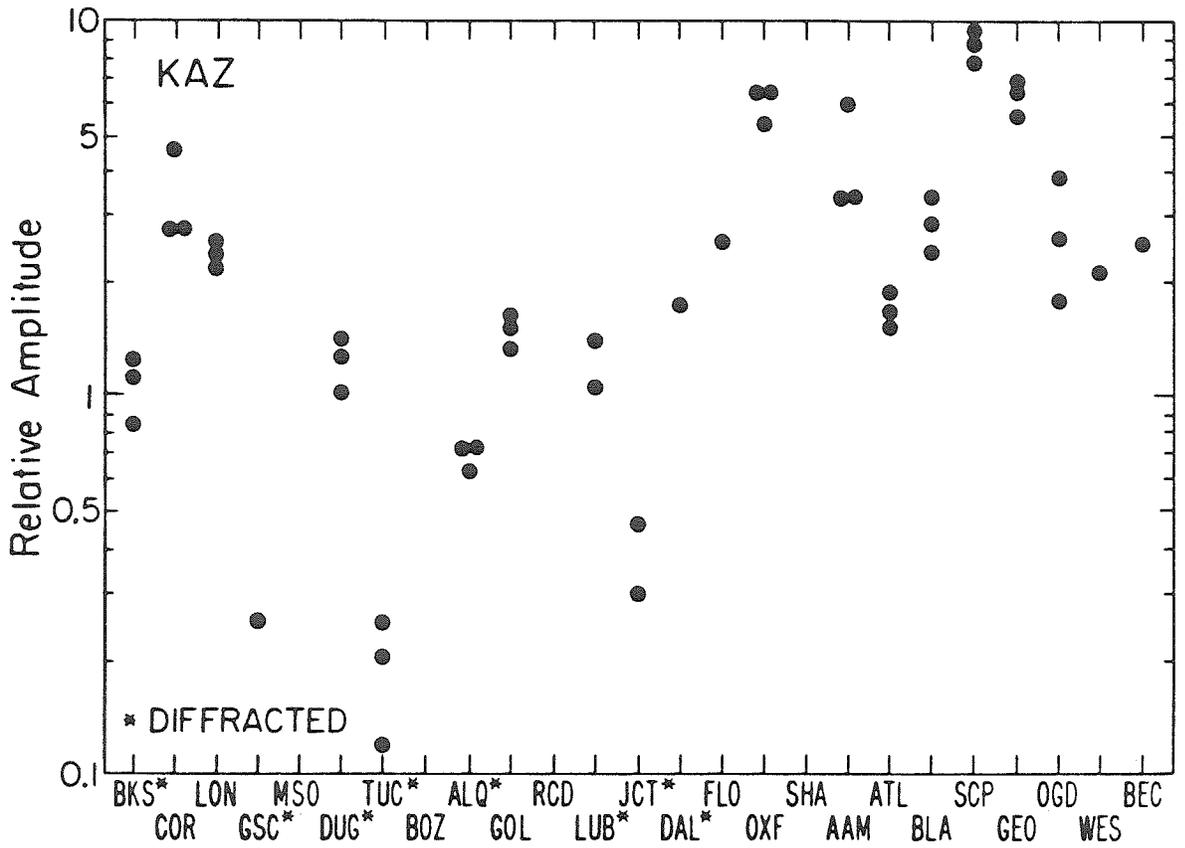
Figure 1.5. The relative AB amplitudes recorded at the WWSSN stations for the three events at the Kazakh site. Notice the reduced amplitudes at stations in the shadow.

depend mostly on the source time function, the mantle attenuation, plus modifications due to the earth response in the D" region. The AB measurement is a good representation of the WWSSN short period amplitude at a dominant period of T~1 sec. Additionally, this measurement is also a reliable indicator of the amplitude decay into the shadow since the first order effect of diffraction is to reduce the overall amplitude, with the relative peak heights changing gradually.

Receiver correction

The primary objective of Butler and Ruff was to determine the short period amplitude behavior of the North American WWSSN stations from the northern azimuth. Using underground nuclear explosions, it was demonstrated that short period amplitudes across the U.S.A. form a consistent pattern from any one test site. Figures 1.4 and 1.5 display the data relevant to the lowermost mantle. The Semipalatinsk site is actually two sites separated by ~70 km, and the agreement between Semipalatinsk East and West indicates that the amplitude variation observed across the U.S.A. is not due to local source structure. In fact, the amplitude patterns obtained for all of the Russian test sites are quite similar (Figure 1.6). Figure 1.6 is important for two reasons: (1) it demonstrates that there are systematic amplitude differences between WWSSN stations, i.e., receiver effects, and (2) there is a factor of ~ 2 scatter about the averaged relative receiver values when comparing the various test sites. Therefore, we can correct
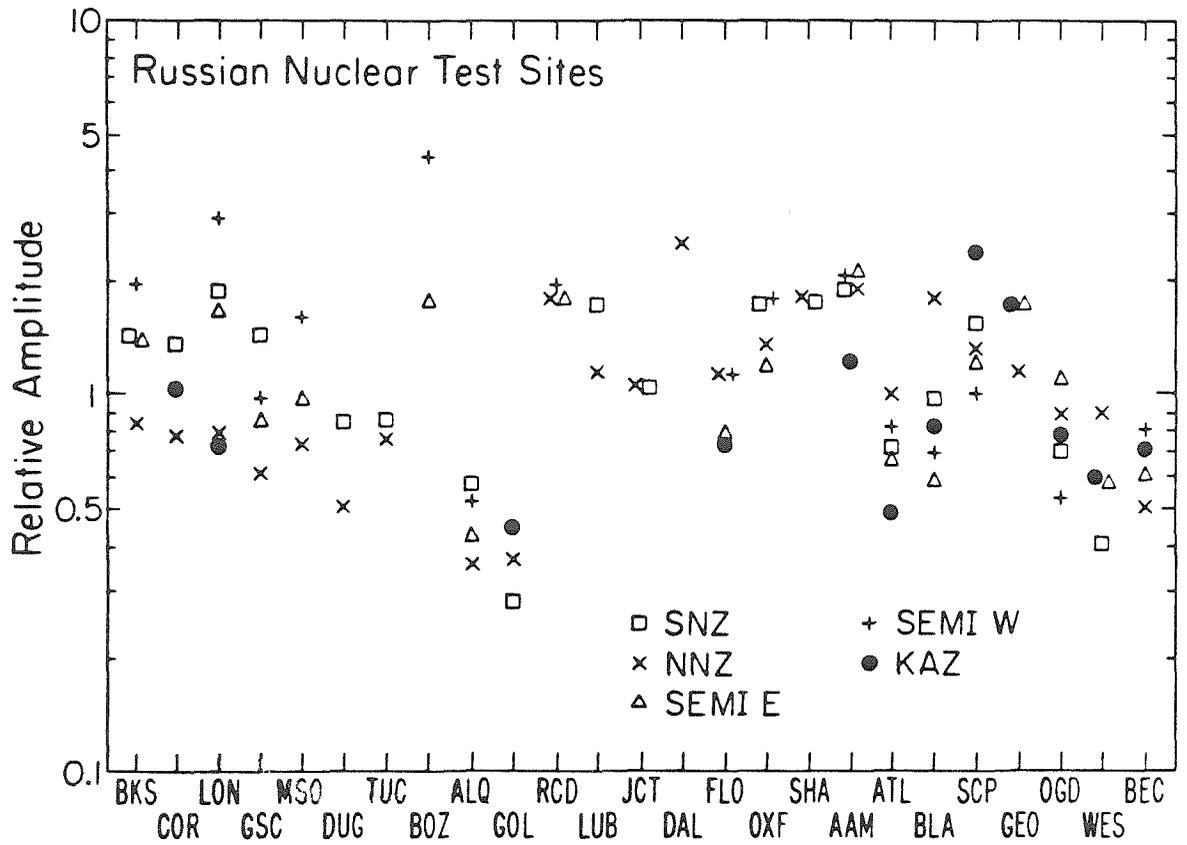
Figure 1.6. Characteristic station amplitudes for the northern azimuth. For each test site, the mean AB amplitude values for all the stations are calculated. Then, the mean values at the five test sites are treated as individual events, and the patterns are scaled to minimize scatter. The diffracted stations and the GOL and DUG values from Semipalatinsk are suppressed in this calculation.

observed short period amplitudes by using the averaged relative receiver values from Figure 1.6, but we should expect the resultant normalized amplitudes to show a scatter of up to a factor of 2. Notice however, that a scatter of a factor of 2 is considerably better than the factor of 6 to 8 if the short period amplitudes were used without any receiver correction.

The receiver correction is applied by first calculating the mean and standard error of the Semipalatinsk data at each station. These averaged amplitude values are then normalized by the average station value obtained from Figure 1.6. The normalized values are then plotted as a function of distance. The same procedure is also applied to the data recorded from the Kazakh site. The details of the normalization are as follows: the baselines of the two amplitude patterns (All-Russian and either Semipalatinsk or Kazakh) are adjusted until the mismatch between the two patterns is minimized in a least squares sense. The stations that are beyond $\Delta = 95$ are suppressed in this calculation, as are the GOL and DUG values from Semipalatinsk due to their highly anomalous amplitudes. Then, the Semipalatinsk (or Kazakh) values are divided by the All-Russian values. Thus, the stations in the non-diffracted range fix the level of the amplitude profile. The agreement between the shadow amplitude values of Semipalatinsk and Kazakh indicates that the two regions sampled have similar properties.

Amplitude versus distance

The corrected amplitudes are plotted versus distance in Figure 1.7. An obvious feature is that beyond $\Delta=95.5$ the amplitudes decrease sharply and steadily from the established baseline. Therefore, as a purely observational conclusion, the apparent core shadow at $T \sim 1$ sec, for the CMB sampled, is at $\Delta=95.5\pm0.5$. This result supports Sacks's (1966) estimate of $\Delta=96\pm1$. The data of the previous studies of Sacks (1966) and Carpenter et al. (1967) are plotted with our results in Figure 1.8. Sacks's study used just two earthquakes and only serves to distinguish the apparent shadow boundary. Carpenter et al. used nuclear events recorded by WWSSN stations, and although we have used approximately twice as many events in total, probably the main reason for the reduced scatter in our profile is the specific correction for the receiver effects from a northern azimuth.

Other features to be noted in the data are: (1) the amplitude decay into the shadow is bracketed quite well. This will be quantitatively discussed later. (2) The pronounced high amplitudes consistently recorded by GOL and DUG from Semipalatinsk. These high values are discussed in Ruff and Butler (1980). (3) An interesting feature in Figure 1.7 is the downward trend in amplitudes from $\Delta\sim80$ to $\Delta\sim90$, followed by an increase to $\Delta\sim95$. One explanation that appears obvious is geometric spreading on a sphere, which causes an amplitude low at $\Delta=90$. This is not the explanation however, as geometric spreading would change the amplitudes by only $\sim1\%$ over the relevant
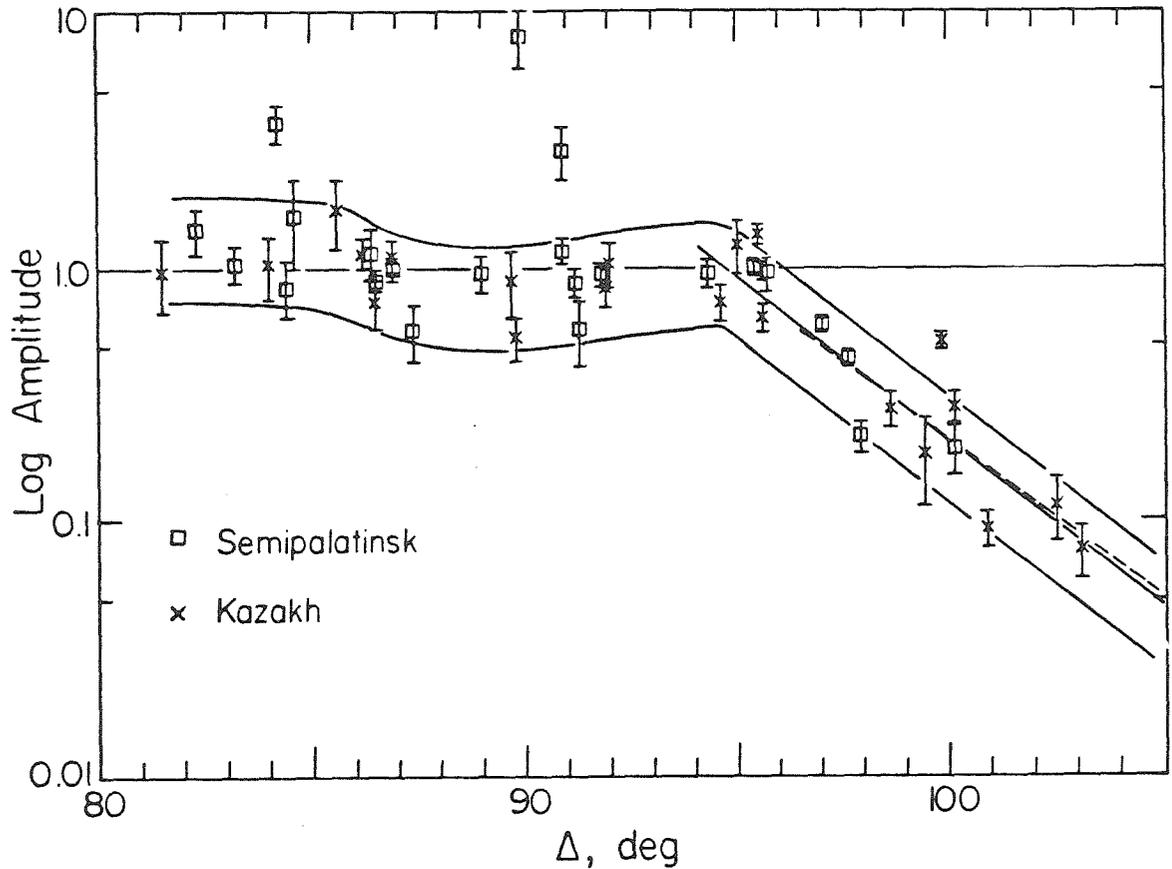
Figure 1.7. Log$_{10}$ amplitude versus distance for the corrected amplitudes. The mean station amplitudes for Semipalatinsk (and Kazakh) are divided by the characteristic station amplitude given by Figure 1.6. This corrects the amplitudes for the common receiver and source effects. In the shadow, the solid line is the best fit line for all points beyond $\Delta = 96$. The three Semipalatinsk points above the data band are COR, DUG, and GOL. The high values at DUG and GOL could be due to lower mantle heterogeneity (see Ruff and Butler, 1980).
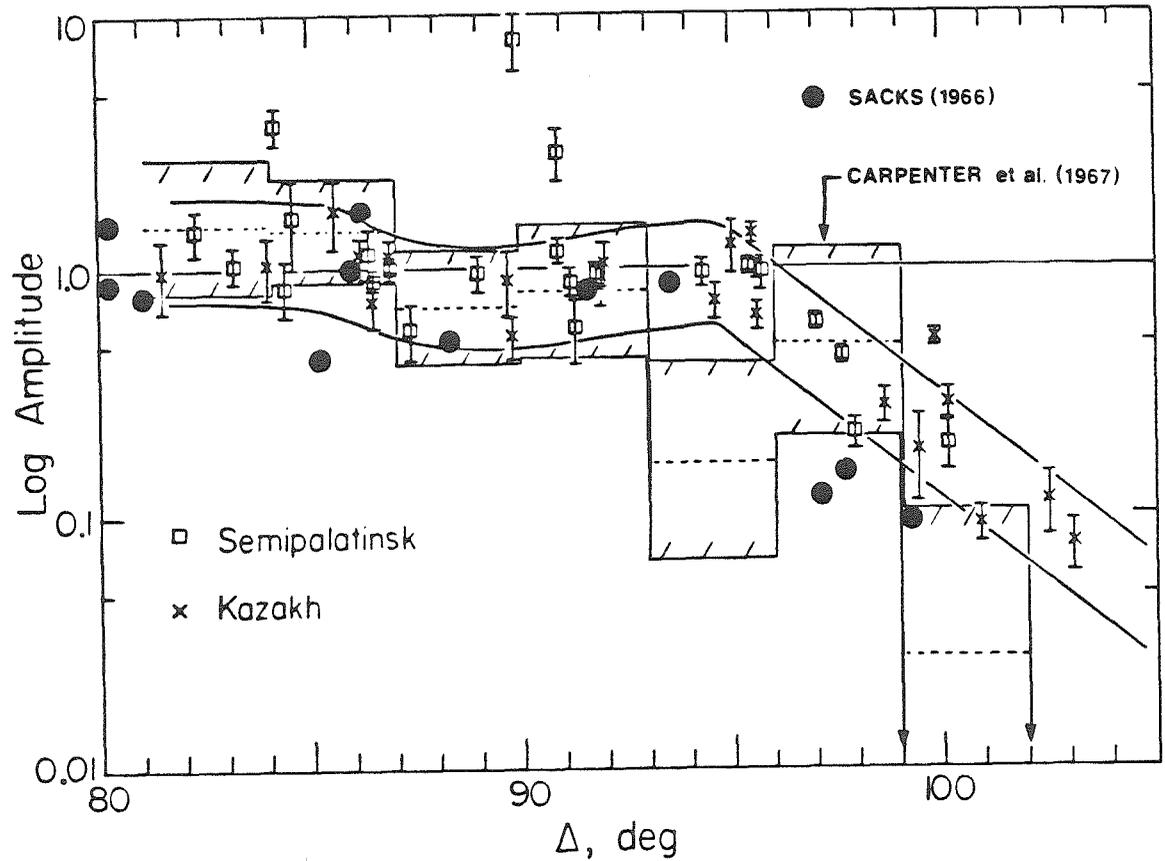
Figure 1.8. Short period amplitudes versus distance of two previous studies are plotted with our data for comparison. Carpenter et al. grouped the WWSSN and Canadian data in 3 deg. blocks. Sacks used WWSSN short period stations and two earthquakes.

distance range. A factor of 1.01 is insignificant in Figure 1.7. This amplitude behavior could be caused by earth structure in D", and we will return to this possibility later. We should note that both source regions, Semipalatinsk and Kazakh, show this amplitude low at $\Delta \sim 90$.

1.3. Interpretation

The main features of the amplitude profile that should be modeled are: (1) the location of the shadow boundary, (2) the slope of the amplitude decay, and (3) the amplitude low at $\Delta \approx 90$ if this proves to be indicative of D" structure. The procedure followed is to construct a synthetic amplitude profile for a smooth earth model that meets the known constraints on D" structure. As previously discussed, the major constraint is that the P wave velocity should not exceed a value of ~ 13.7 km/s in D". A smooth earth model satisfying this condition is 1066B (Gilbert and Dziewonski, 1975). This model, based on normal mode inversion, has a nearly contant P wave velocity in D". As discussed in the following section, a major conclusion is that the observed amplitude profile has a resolvably faster decay of amplitudes than predicted by 1066B. Therefore, there is an extra "attenuation" required in the shadow zone, and we will consider four possible explanations, (1) a smooth high velocity gradient, (2) scattering, (3) a thin low-Q zone at the CMB, and (4) a rough velocity structure. First, we discuss the synthetic amplitude profile for 1066B.

Synthetic seismograms

Two techniques for constructing body wave synthetic seismograms are used in this paper, the full wave method and the generalized ray method. These methods have been described elsewhere (see references in Appendix A), so we will not discuss any details here. The relative merits of these methods can be briefly summarized as follows. With the full wave technique, the interaction of the wavefront with a spherical boundary into the shadow zone presents no special difficulties, though there are currently restrictions on the velocity models that can be considered. With the generalized ray method, any velocity model can be used, however, for long diffracted waves (or waves trapped in a waveguide) one must be careful to include the necessary multiple rays. The generalized ray method has been successfully applied to many upper mantle and crustal problems where it works quite well. The full wave method is better suited for problems dealing with core phases with a regular velocity structure.

The synthetic amplitude profiles are obtained by convolving the earth response with the instrument, an appropriate source function, and attenuation operator; then measuring the resultant AB amplitudes. Figure 1.9 shows the results for 1066B computed with the full wave method. The synthetic AB values are within the data band well into the shadow, but that is mostly due to the fact that the 1066B points start to decay at $\Delta\sim92$ instead of at $\Delta\sim95$, and it is visually apparent that the amplitude decay of the synthetics is less than the observed decay.
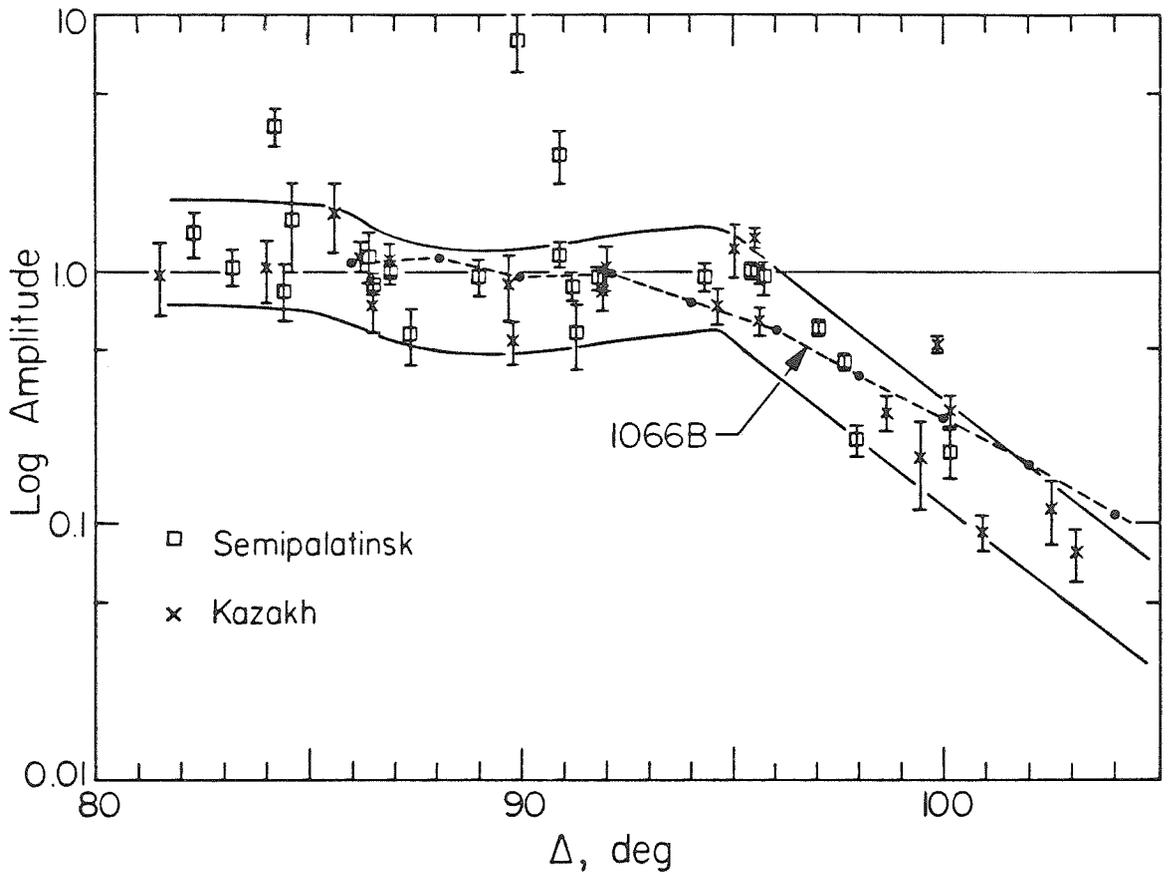
Figure 1.9. The synthetic AB amplitude profile for the 1066B earth model, calculated with the full wave method. The data are the same as Figure 1.7. In the shadow, the slope of the data band (-1.28 log/10 deg.) corresponds to an average best fit line through the data.

To quantify this comparison, we compare the slopes of the amplitude decay in the units of $\log_{10}/10$ deg. The decay slope of 1066B is -0.92. For the data: the decay slope including all the points beyond $\Delta=94$ is $-1.32\pm.14$ (standard error). If we take the values beyond $\Delta=96.0$, then the decay slope is $-1.25\pm.22$). Therefore, we tentatively conclude that there is a resolvable difference between the amplitude decay of 1066B and the data. As an independent check, the synthetic amplitude decay of the generalized ray synthetics (see Figure 1.13) is -0.88, consistent with the full wave value. Additionally, an asymptotic formula can be used to calculate the decay at a given period for a particular velocity gradient. Using the result given in Chapman and Phinney (1972), the asymptotic slope for $T=1$ sec is -0.91, which agrees with the value obtained from the synthetics. Therefore, we conclude that the amplitude decay slope of 1066B does not match the observed amplitude decay. Also, the 1066B synthetics do not produce the amplitude low at $\Delta\sim90$, nor the sharp shadow boundary.

It is important to realize that a faster amplitude decay is associated with a steeper positive (increasing with depth) velocity gradient. Figure 1.9 implies that a velocity gradient steeper than that of 1066B is required to satisfy the data. This implied velocity gradient change may violate $dT/d\Delta$ data. We will now consider the effect of a steeper velocity gradient.

Steep positive velocity gradient

Though the asymptotic formula could be used to estimate the velocity gradient implied by the observed amplitude decay, for completeness we have calculated the synthetic AB profile for an earth model with a steep positive gradient in D". PEM (Dziewonski et al., 1975) is an earth model which describes the physical properties as a first or second order polynomial as a function of radius. It is a very smooth model, and consequently does not satisfy travel time data in parts of the mantle, including D" where the velocity gradient in the lower mantle is continued through D". This model demonstrates the effect of a steep gradient in D". The AB profile computed for PEM is plotted with the data in Figure 1.10. The steep velocity gradient of PEM results in a satisfactory decay slope and also produces the sharp shadow boundary, but it is clearly not consistent with the amplitude data due to the ~2 deg. shift in the shadow boundary. Additionally, the dT/dΔ minimum is 4.42, and the smooth velocity gradient does not cause the abrupt change in dT/dΔ at Δ~90.

Reducing the core radius would move the PEM shadow boundary to larger distances. For the velocity gradient of PEM, a reduction in the core radius of ~ 40 km moves the shadow boundary by only 1 deg. Estimates of the core radius from PcP and normal mode observations allow only a ~ 10 km change. Also, reducing the core radius further exaggerates the dT/dΔ discrepancy.

Thus, although the amplitude decay seems to imply a velocity model
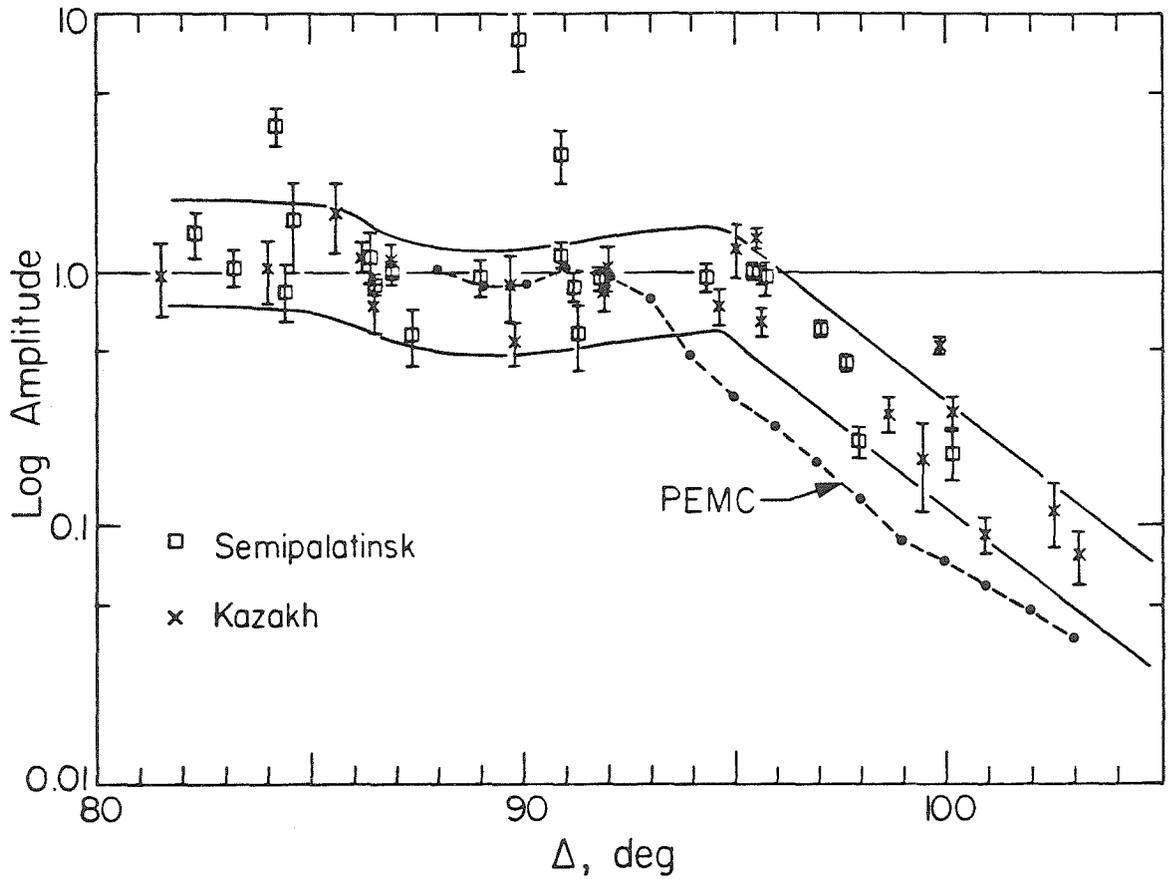
Figure 1.10. The synthetic AB amplitude profile for the PEMC model, calculated with the full wave method. The decay slope of PEMC is acceptable, but the shadow boundary shift is not.

with a steep positive gradient, such a model is inconsistent with the amplitude data due to the associated shadow boundary shift. To retain a smooth velocity model such as 1066B for D", we must find some process to sharply decrease the amplitudes in the shadow. Two processes, scattering by small scale inhomogeneities and anelastic attenuation will be considered next.

Scattering

Scattering by small scale inhomogeneities in D" has been proposed as an explanation for the precursors to PKIKP (e.g., Doornbos and Husebye, 1972; Haddon and Cleary, 1973; King et al., 1973; Husebye et al., 1976). Though it is now generally agreed that the precursors do represent scattered energy, a more detailed analysis has suggested that the scatterer has a coherent structure over a length scale greater than 100 km (Haddon, 1978). A scattered arrival is characterized by precursory energy building up to a maximum amplitude, then slowly decreasing. The wave envelope is typically the most distinctive feature of scattered arrivals. This behavior has been reproduced for the PKIKP precursors by using a first order scattering theory, in which the scattered energy is approximately 1% of the incident energy (e.g., Doornbos, 1978). If there are scatterers in the lower part of D", forward scattering will occur and we would expect to see the scattered energy in the seismograms. Given the inherent scatter of a factor of ~ 2 in the amplitude profile, an additional 1% scatter is negligible. The mismatch of the data and the 1066B synthetics (Figure 1.9) require at

least a 50% amplitude decrease. If the scattering is strong enough to decrease the AB amplitude by 50%, then the seismograms should have a scattered appearance, i.e., a wave envelope with precursory arrivals.

As seen in Figure 1.3, the arrivals do not have this appearance and, as previously discussed, we can easily model the first part of the seismograms without introducing extraneous arrivals. The scattered energy could be focused to interfere destructively at a particular distance, but it could not explain the diminished amplitudes at all ranges greater than 95 deg. Therefore, we conclude that forward scattering is not important with regard to the amplitude decay mismatch.

Another scattering process would be associated with a rough CMB, where the scattered energy might go into the core so that we would observe "clean" seismograms. In this case, the amplitude loss would accumulate and might approach the required amplitude deficit. However, we present a heuristic argument against this possibility. The maximum allowable topography of the CMB is about ~ 1 km (seismologically and mechanically). If we take the seismic energy density to be nearly uniform throughout D", then the energy lost by the rough CMB can be considered equivalent to moving the CMB up by 1 km. By assuming an extremal situation in which the scattered energy destructively interferes with seismic energy about it, then the CMB is effectively moved up by 2 km. Changing the core radius by 2 km has a negligible effect on the shadow boundary, and consequently would not cause the required amplitude reduction. Depending upon the length scale of the CMB undulations, the amplitude reduction due to scattering might be

concentrated at a particular distance, but it is unlikely to cause a substantial amplitude reduction over a range of 10 deg.

Therefore, while it is difficult to fully assess the role of scattering, heuristic arguments indicate that it is not important. Forward scattering of ~ 1% would not explain the amplitude mismatch. A stronger forward scattering is not allowed by the waveform coherence. The other possibility, scattering into the core by a rough CMB, does not seem to be important as the CMB topography is not large enough to reduce the amplitudes by 50% over a large distance range.

Low-Q zone

An obvious process to reduce amplitudes in the shadow is anelastic attentuation due to a low-Q layer in D". To evaluate this process, there are two levels of consideration. First, the effects of attenuation can be considered "geometrically". Body wave attenuation is usually parameterized by the variable $t*$, $t*=$(travel time)/(average Q). The attenuation is then given by the factor $\exp(- \pi f t*)$, with f the frequency. The standard procedure for calculating the body wave $t*$ is to sum the contributions from each layer, $t*=2\Sigma \, dt*_i$ where $dt*_i$ is the $t*$ for a particular layer, and the sum is taken over the total number of layers. The attenuation of diffracted waves can be accommodated geometrically by adding a $dt*_{CMB} = $ (travel time)$_{CMB}/Q_{CMB}$. This method is satisfactory for moderate gradients in Q^{-1}. If a thin low-Q layer is used, this geometric method will not give accurate results.

As a first approximation, the geometric method is used to estimate the Q value required to adjust the synthetic amplitudes. The 1066B amplitude points are already lower than the data before $\Delta=95$, therefore a low-Q zone should be confined to the lower part of D" so that it primarily affects the amplitudes beyond $\Delta=95$. Considering various Q distributions, it was found that any low-Q layer thicker than ~60 km significantly affects the t* before the shadow as the seismic rays turn around in the low-Q layer. In all cases there is never a sharp amplitude decay such as the observed shadow boundary break. Adjusting the Q value at the CMB, it was determined that a layer 60 km thick at the base of D" with a $Q_\alpha=200$ gives a satisfactory value for dt^*_{CMB}. The Q_α above this layer is high, ~ 1000.

With the above Q model, the seismic rays are travelling a large distance near a sharp contrast in $\bar{Q}^1$. While the "geometric attenuation" uses one t* operator for the entire seismogram, a better approximation is to use a different t* operator for different parts of the seismogram. In the context of generalized ray theory, this occurs because the rays which travel in the low-Q layer are more attenuated than the rays above the low-Q zone. Thus, the apparent t* will be different for each generalized ray and there will be rays arriving at the same time but with different t*. This type of attenuation approximation is derived in Appendix A. Note the approximations implied by this treatment: (1) attenuation does not affect the reflection coefficient, and (2) there is a second order phase effect that is ignored. Anelasticity can effect the reflection coefficient, but is

important only for critical incidence with a very low Q, ~ 10-20 (Borchedt, 1977; Krebes and Hron, 1980). The second order phase effect occurs when p has a small imaginary part, of particular importance for the head waves. Thus, with a Q_α=200 we expect the anelastic calculations to be accurate, except when the synthetics are mostly comprised of head waves that have propagated a long distance in layers with a large contrast in Q^{-1}. We have modeled the low-Q zone as a jump in Q, with Q constant above and below the discontinuity. Although a steep gradient in Q across the thin zone might seem more appropriate, the steep gradient structure cannot be resolved from a thin constant Q layer.

The synthetic AB profile for the anelastic model LMQ1 is shown in Figure 1.11. The LMQ1 model has a low-Q layer with Q_α=200 in the lowermost 60 km and Q_α=800 above (note: Q_α is probably higher than this, but increasing the value will not seriously affect the results), and the elastic structure is 1066B.

In Figure 1.11, the amplitude profile for 1066B with a high Q throughout D" (t*=constant) is plotted for comparison. The amplitudes for LMQ1 depart from those of 1066B at about Δ~87. The slope of the amplitude decay in the shadow is not seriously affected by the low-Q layer, as most of the attenuation takes place before the shadow when the rays first enter the low-Q layer. More extreme Q models were considered, but in no case could the decay slope be significantly altered without prohibitive effects on waveshapes. Also, in general, a low-Q layer causes a smooth decline in amplitudes well before the shadow
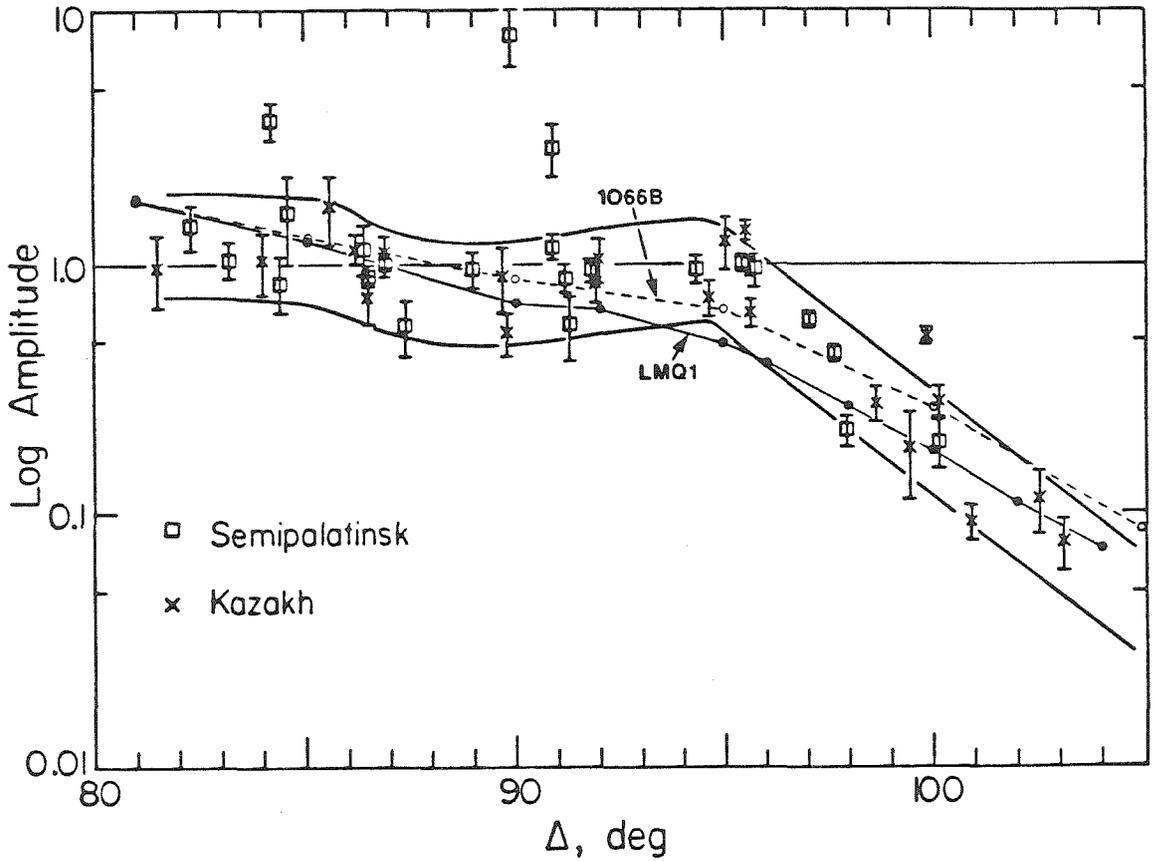
Figure 1.11. The synthetic AB amplitude profile for the anelastic model LMQ1, calculated with the anelastic generalized ray method. LMQ1 has a 1066B elastic structure, with a 60 km low-Q (Q_α = 200) layer at the base of D". The amplitude profile for 1066B (with high Q) calculated with the generalized ray method is also plotted so the effects of the low-Q layer can be seen.

boundary. Therefore, a low-Q layer cannot explain the sharper features of the amplitude profile, i.e., the amplitude minimum at $\Delta=90$ and the distinct shadow break.

To summarize our interpretation thus far, we concluded that 1066B (or any smooth velocity model) is inconsistent with the observations. To reduce the amplitudes two processes, scattering and anelastic attenuation, are considered. Scattering is an unlikely explanation due to the large amplitude reduction required and the coherence of the seismograms. Though the synthetic amplitude profile for the anelastic model LMQ1 almost fits within the data band, it does not satisfy the data in that the decay slope is not consistent and no sharp features can be produced with a low-Q layer. Thus far, we have assumed a smooth elastic structure. Since the observed amplitude proflile has not yet been satisfactorily explained, this assumption should be rescinded. In the next section, we consider rough models.

Layered velocity structure

If we allow a more complicated velocity structure, then interference effects might produce sharp features in an amplitude profile. Constructing a velocity model that gives the desired amplitude behavior is quite simple. The amplitude decay in the shadow requires a steep positive velocity gradient. Due to $dT/d\Delta$ constraints, the velocity value at the CMB is fixed to be less than ~ 13.68 km/sec. A steep velocity gradient (such as that of PEM) then prescribes the

velocity in the lower part of D". As the velocities at the CMB and the top of D" are nearly equal, allowing a steep gradient in the lower part of D" introduces a low velocity channel in the upper part of D". At ranges where the geometric ray is bottoming near the low velocity channel there will be two seismic arrivals, from above and below the channel. These two arrivals may interfere constructively and destructively, dependent upon the distance and the dominant period of the observed arrival. The presence of the low velocity channel also shifts the shadow boundary to a larger distance. Thus, a steep gradient in the lower D" and the low velocity channel might be able to explain the shadow decay and the sharper features of the profile.

POLAR1 is a model constructed in the manner described above (see Figure 1.12), and the synthetic amplitude profile is shown in Figure 1.13. The synthetic amplitudes behave in the desired manner: low at Δ~90-92, enhanced at Δ~95, followed by a sharp decline into the shadow with the appropriate decay slope. The amplitudes decline from Δ=87 to Δ=92 due to the low velocity channel, i.e., diffracting along the "high velocity lid". The amplitude enhancement from Δ=92 to Δ= 95 is caused by the constructive interference of the arrivals from above and below the low velocity channel. Beyond Δ=95, the arrival from the lowermost D" is the first arrival. Many D" velocity models were tested, and Figure 1.13 also show the results for two other successful models, POLAR6 and POLAR9. Both of these models display the same basic amplitude behavior as POLAR1.

Only the first order rays were used in computing the synthetic
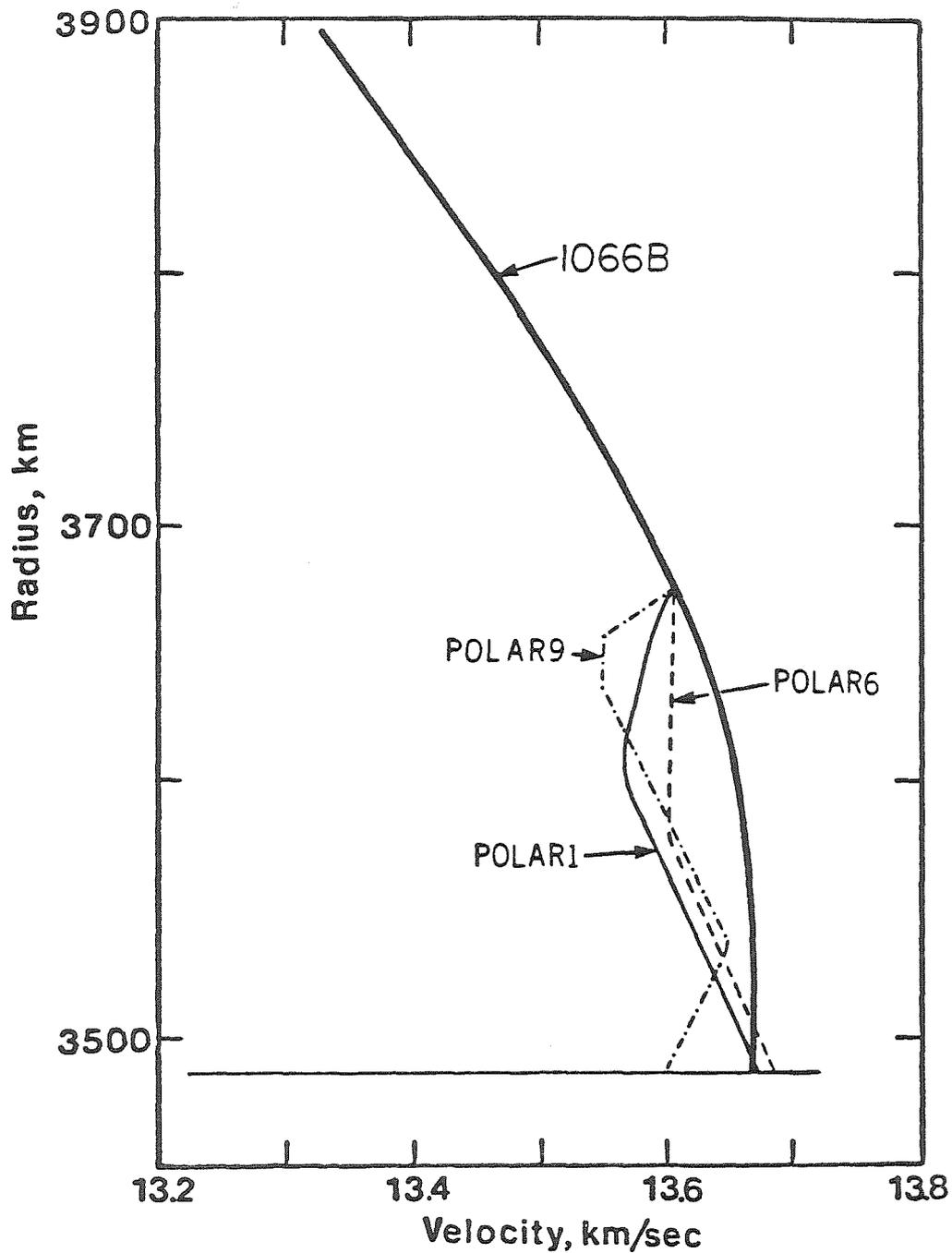
Figure 1.12. Three of the POLAR velocity models that explain the amplitude data. The 1066B velocity structure above D" is used in all cases. The Q is assumed to be high throughout D".
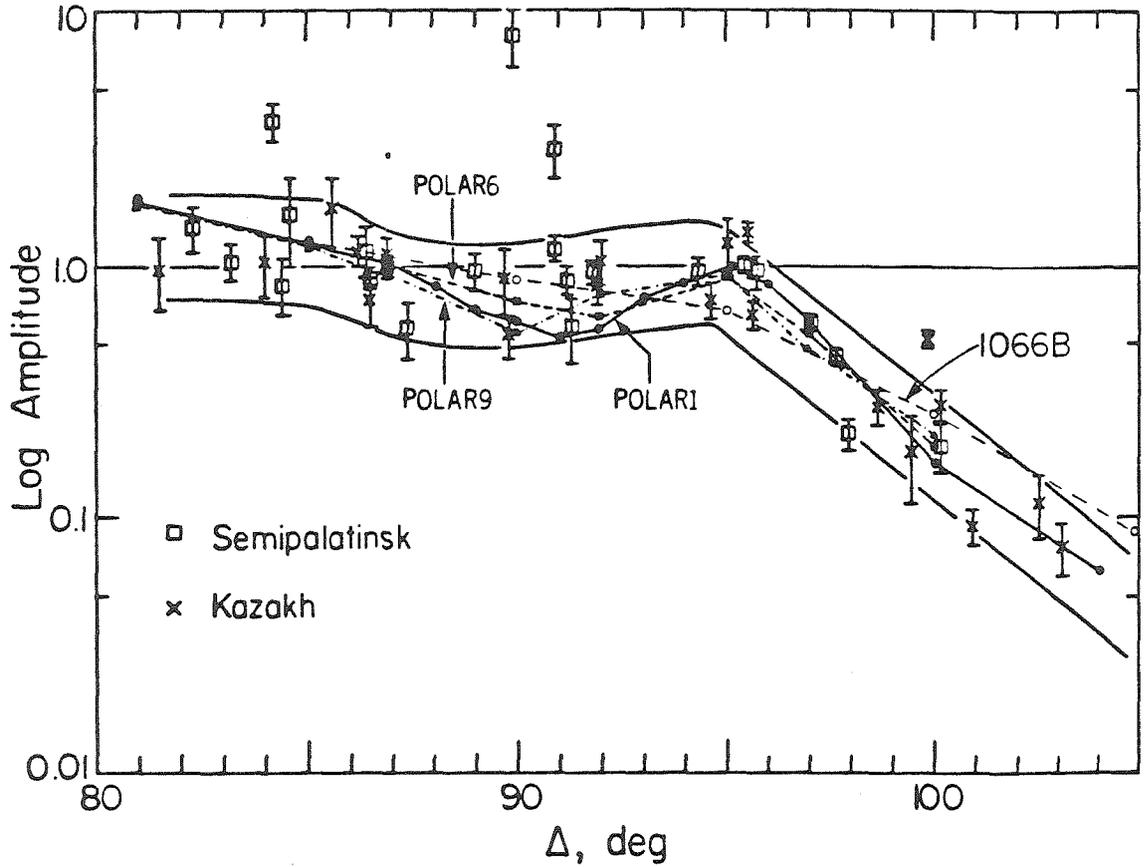
Figure 1.13. The synthetic AB amplitude profiles for the POLAR models, calculated with the generalized ray method. The 1066B model is shown for comparison. Notice that the POLAR amplitudes oscillate about the smooth 1066B profile; and the sharp decay into the shadow is in better agreement with the data than the 1066B amplitudes.

amplitude profiles. Caution must be used in extending the profiles far into the shadow in the presence of a low velocity channel as the amplitudes of the first order rays decrease relative to the multiple reflected within the thin layers in the lowermost mantle (see Appendix A). Sets of multiple rays have been computed, and while the multiples may influence the amplitudes beyond $\Delta \sim 100$, the amplitude behavior from $\Delta \sim 80$ through $\Delta \sim 97$ is largely unaffected.

It is perhaps surprising that such subtle changes in the D'' velocity structure can modify the amplitudes by up to a factor of $\sim$ 2. A brief examination of Figure 1.13 reveals that the amplitudes of the POLAR models "oscillate" about the smooth profile of 1066B. The low velocity channel bifurcates the seismic energy into two arrivals over a small distance range, delaying the arrival from the lowermost D''. The extent of the low velocity channel determines the location of the amplitude "oscillations". The three models shown in Figure 1.12 and their amplitude profiles in Figure 1.13 demonstrate the sensitivity of the interference effect. The POLAR6 model has the least developed low velocity channel, actually just a constant velocity zone, and consequently has the smallest amplitude oscillations about 1066B. There is also a constraint on the extent of the low velocity channel. For example, one unsuccessful model had a 1% decrease in velocity (to 13.47 km/sec) at the top of D'' immediately followed by a steep gradient, and the amplitude enhancement occurred at $\Delta \sim 97$ instead of $\Delta \sim 95$.

The observed amplitude profile is sensitive to interference effects because: (1) the observations are at short periods (AB time is ~ 0.5

sec), (2) the dominant period of the observations is nearly the same throughout the profile. Otherwise, with observations of widely varying dominant period, the interference effect would be smoothed out. As the two distinct seismic arrivals are never separated by more than 0.7 sec, long period seismograms would be sensitive only to the averaged velocity gradient in D". Therefore, long period amplitude studies cannot resolve the velocity structure of the POLAR models.

An important check on the POLAR models is whether the interference has an observable effect on the waveforms. Figure 1.14 shows the computed synthetic seismograms for the POLAR models and 1066B. Relative to 1066B, consistent differences are that the second peak moves forward in time and the peak to peak amplitudes change slightly at $\Delta \sim 95$. However, the scatter in the observed waveforms before $\Delta \sim 87$ is nearly as large as the interference induced waveform changes (Figure 1.15). Thus, the POLAR models do not cause a resolvable waveform distortion. If receiver response functions could be deconvolved from the observed waveforms, then the D" interference effect may be observed.

The interference effect is apparent in $dT/d\Delta$ measurements. Figure 1.16 shows the $dT/d\Delta$ for POLAR1 measured in three ways: (1) geometric arrival time, (2) the times of the first peak, and (3) the times of the second positive peak. While the low velocity channel is revealed by the geometric arrival times, in practice one cannot measure the true first arrivals to determine $dT/d\Delta$, and various techniques for measuring $dT/d\Delta$ have been used, including waveform correlation. The interference effect causes oscillations in $dT/d\Delta$ for different parts of the seismograms.

184
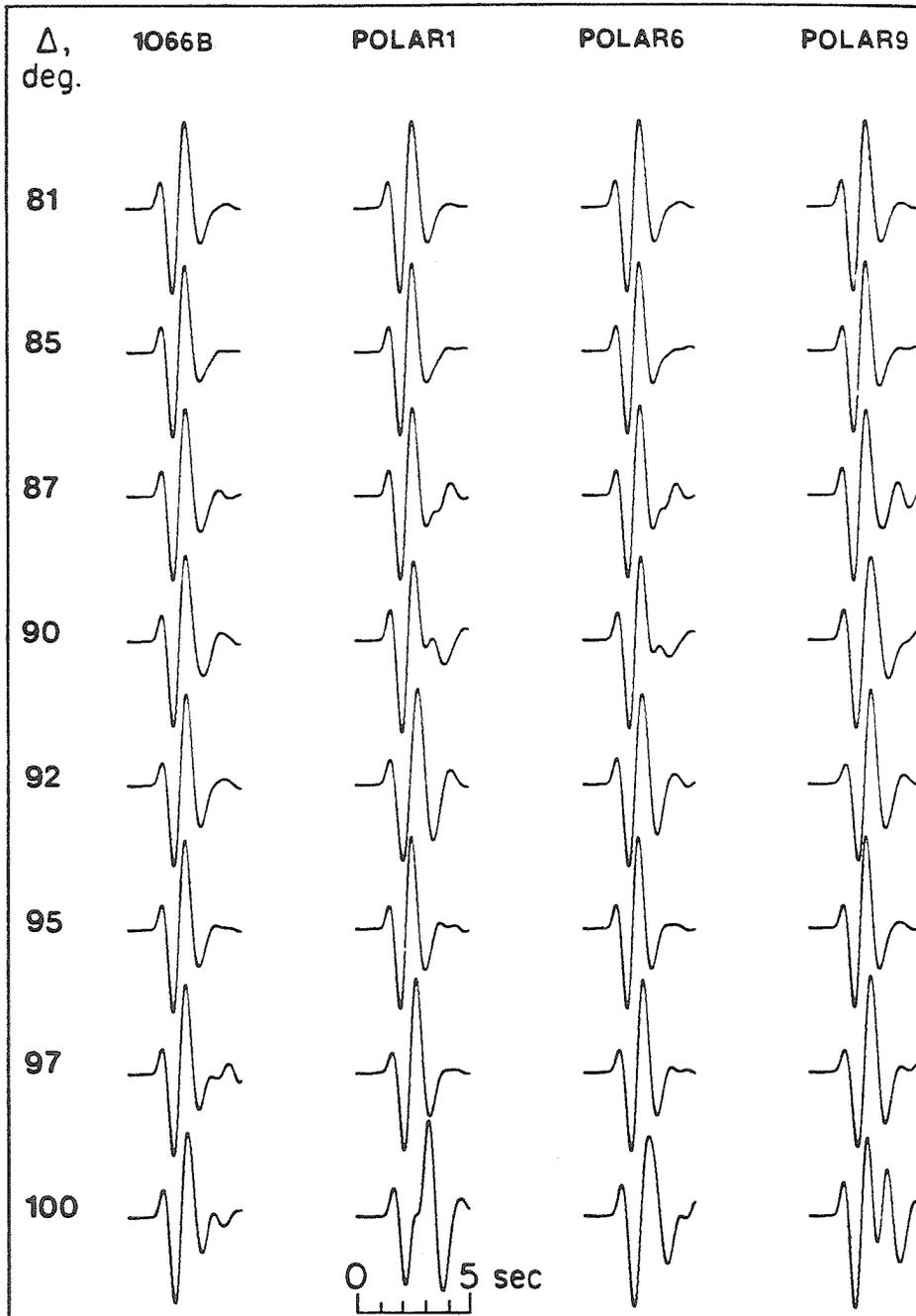
Figure 1.14. The synthetic WWSSN short period seismograms for the POLAR models and 1066B. The peak to peak amplitudes for all seismograms are the same. The synthetics were computed with a Von Seggern and Blanford (1972) explosion source (k = 5.0, B = 2.0, pP-P = 0.5 sec.) and t* = 1.0. There are other combinations of the source parameters and t* which match the observed waveforms and AB amplitudes. The PcP phase is quite small at these distances and not observable as a distinct arrival (see, e.g., Kanamori, 1967).
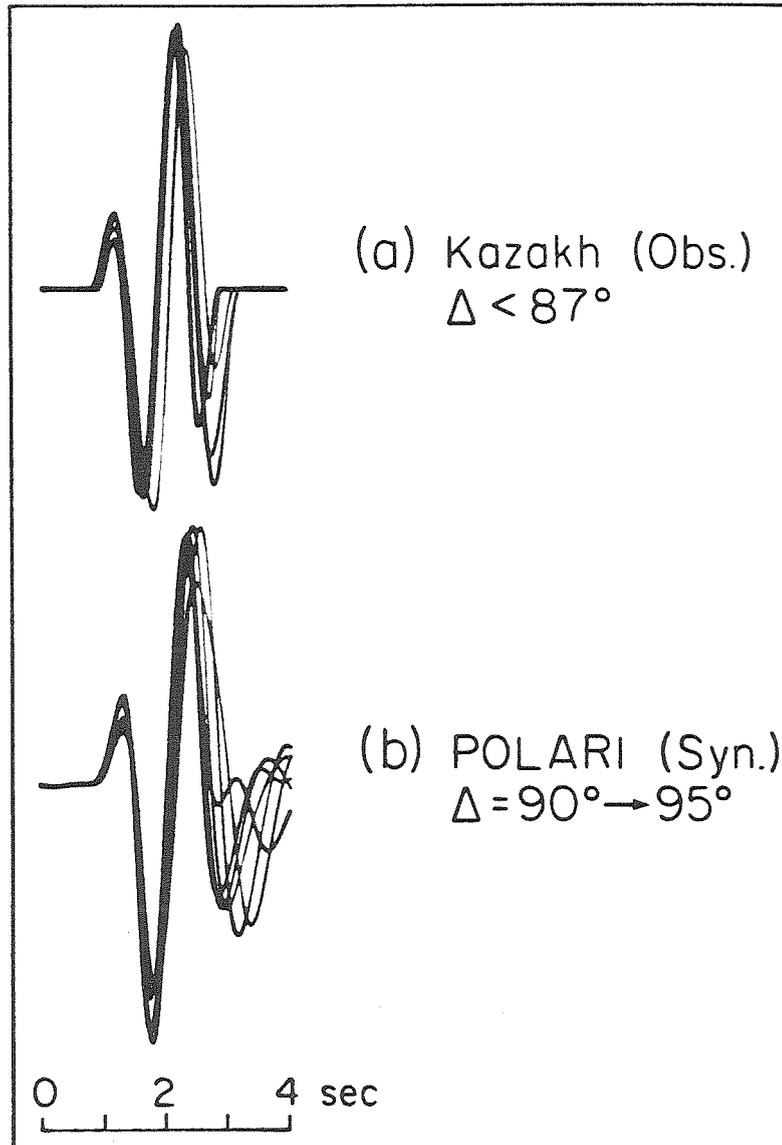
Figure 1.15. The scatter in the observed waveforms before Δ = 87 is
shown in (a).   The seismograms of two stations (SCP and GEO) from the
three Kazakh events are lined up on the first peak and the peak to peak
amplitudes are the same.   In (b), the synthetic seismograms for POLAR1
(from Figure 1.14) are plotted in the same manner.   The interference
effects in the range Δ = 90 to 95 are apparent in the synthetics, but
the intrinsic waveform scatter at two relatively "clean" stations is
nearly as large.
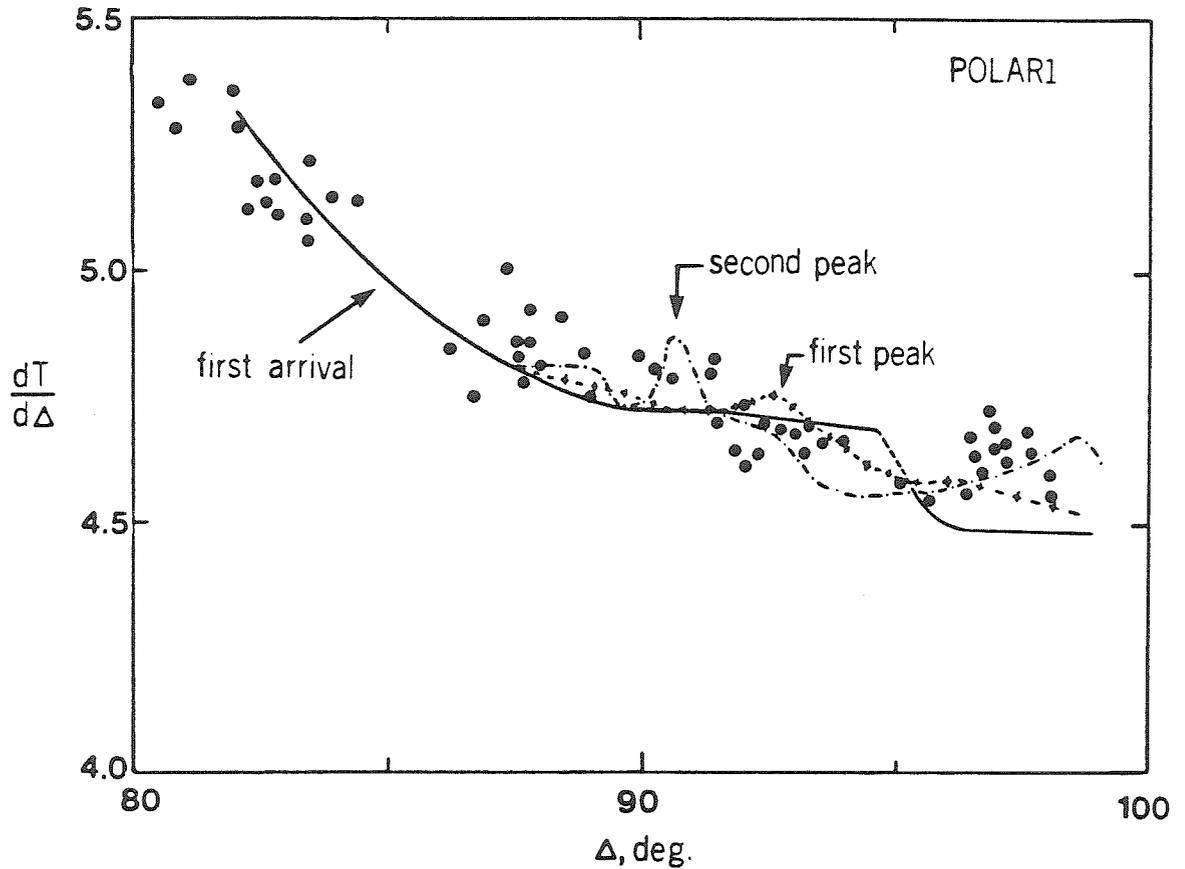
Figure 1.16. The dT/dΔ curve for POLAR1 calculated in three ways. the
geometric arrival times, times of the first peak, and times of the
second peak (the synthetics in Figure 1. 14 are used). The data are
from Chinnery (1969). The interference effects are reflected in the
dT/dΔ of the peaks. A highly accurate small aperture seismic array is
required to isolate the dT/dΔ oscillations. Otherwise, the interference
effects show up as "scatter" in dT/dΔ.

Incoming signals with different dominant periods have different patterns

of dT/dΔ oscillation. Thus, the observed scatter in dT/dΔ measurements

beyond Δ=90 could be explained by the POLAR models. The combination of

different measuring techniques, varying period of the incoming signal,

and aperture of the seismic array would cause an apparent scatter in

dT/dΔ. Therefore, the POLAR models not only explain the short period

amplitude profile, but these models may also explain the dT/dΔ results.

1.4 Conclusion

An important conclusion of this paper is that the short period

amplitudes are not consistent with the usually assumed smooth velocity

models in D". This conclusion is based on two aspects of the amplitude

profile: the decay slope in the shadow and the presence of an

"oscillation" in the amplitude profile before the shadow. This

amplitude oscillation is related to D" structure as the distances to the

stations are different for the two test sites (see Table 1.1), and the

sharp amplitude decline at the shadow boundary and the amplitude low at

Δ~90 occur in both data sets, Semipalatinsk and Kazakh.

The LMQ1 anelastic model is not satisfactory for the same reasons

as 1066B: the synthetic profile is too smooth and the decay slope is

not steep enough. The synthetic profile for LMQ1 falls within the data

band, and this might be interpreted as a satisfactory match. However,

we consider the sharper features in the observed amplitude profile to be

indicative of D" structure, and a thin low-Q zone cannot explain these

features. A conclusion regarding anelastic models would be stronger if the amplitude profile could be extended 5 deg. further into the shadow (see Figure 1.11).

The POLAR velocity models fit all of the above-discussed features in the amplitude profile. These rough models are seismologically appealing in that they might explain other results as well: the $dT/d\Delta$ scatter and complications in seismograms beyond $\Delta = 90$ (see Johnson, 1969). The POLAR velocity structure could cause complications in PKP at grazing incidence into the core. Also, a corresponding rough structure for the S wave velocity would affect ScS at particular ranges (see Mitchell and Helmberger, 1973). Thus, a POLAR type velocity model, which is required to explain our amplitude data, may also reconcile various "anomalous" observations concerning D" and the CMB.

The differences in velocity between the POLAR models and a smooth model are rather small, but the geophysical implications are significant. A D" model that has a smoothly decreasing velocity with or without a low-Q zone would be consistent with the D". layer interpreted as a single thermal boundary layer. Admitting a rough velocity model for D" does not allow the above interpretation. A velocity decrease at the top of D" followed by a "normal" velocity gradient below implies a compositional and/or phase change (gradual or distinct) at the top of D". The low velocity channel could be interpreted as a double thermal boundary layer with very thin boundary layers, if there is sufficient heat flux from the core. The POLAR9 model was constructed to be consistent with very thin boundary layers at the top of D" and at the

CMB. Notice that the existence of a thermal boundary layer at the top of D" still implies a compositional stratification. A candidate compositional change is the Ca,Al rich refractory assemblage proposed by Ruff and Anderson (1980), though there are probably other possibilities.

Regarding the Q structure of D", there have been suggestions of a low Q in the lower mantle (e.g., Anderson and Hart, 1978). These suggestions are based on long period data, and with a moving absorption band model it is possible that the attenuation is less at T=1 sec. Though we rejected a thin low-Q zone in the lowermost D" as an explanation of the amplitude profile, a Q of ~ 500 throughout D" could be superposed on the POLAR models.

The limited sampling of the CMB by this study should be noted. This type of study should be used to sample the D" layer in other localities. Although no other geometry of sources and stations is as favorable, the results of the present study indicate diagnostic features in the amplitude profile that might be isolated in other studies. If D" is characterized by a rough velocity model, then the many suggestions of lateral variations might be due to scatter induced by a rough model. Confirming lateral variations in D" would be important, as mapping out the variations could indicate the cause of the complicated velocity strucuture.

References for Part II

Alexander, S. S. and Phinney, R. A., 1966. A study of the core-mantle boundary using P waves diffracted by the Earth's core, J. Geophys. Res., 74, 5943-5958.

Anderson, D. L., 1975. Chemical plumes in the mantle. Geol. Soc. Am. Bull., 86, 1593-1600.

Anderson, D. L. and Hart, R. S., 1978. Attentuation models of the Earth, Phys. Earth Planet. Int., 16, 289-306.

Bolt, B. A., 1972. The density distribution near the base of the mantle and near the Earth's center, Phys. Earth Planet. Int., 5, 301-311.

Borchedt, R. D., 1977. Reflection and refraction of type-II S waves in elastic and anelastic media, Bull. Seis. Soc. Am., 67, 43-67.

Bullen, K. E., 1950. An Earth model based on a compressiblity-pressure hypothesis, Mon. Not. R. astr. Soc. Geophys. Suppl., 6, 50.

Burdick, L. J. and Helmberger, D. V., 1978. The upper mantle P velocity structure of the western United States, J. Geophys. Res. 83, 1699- 1712.

Burdick, L. J. and Helmberger, D. V., 1979. Time functions appropriate for nuclear explosions, Bull. Seism. Soc. Am., 69, 957-973.

Butler, R. and Ruff, L. J., 1980. Teleseismic short-period amplitudes: source and receiver variations, Bull. Seism. Soc. Am., 70, 831-850.

Carpenter, E. W., Marshall, P. D., and Douglas, A., 1967. The amplitude- distance curve for short period teleseismic P-waves,

Geophys. J. R. astr. Soc., 13, 61-70.

Chapman, C. H., 1973. The earth flattening transformation in body wave theory, Geophys. J. R. astr. Soc., 35, 55-70.

Chapman, C. H., 1974a. Generalized ray theory for an inhomogeneous medium, Geophys. J. R. astr. Soc., 36, 673-704.

Chapman, C. H., 1974b. The turning point of elastodynamic waves, Geophys. J. R. astr. Soc., 39, 613-621.

Chapman, C. H. and Phinney, R. A., 1972. Diffracted seismic signals and their numerical solution, Methods in Computation Physics, 12, 165-230, Academic Press, Inc., New York.

Chinnery, M. A., 1969. Velocity anomalies in the lower mantle, Phys. Earth Planet. Int., 2, 1-10.

Choy, G. L., 1977, Theoretical seismograms of core phase calculated by a frequency dependent full wave theory, and their interpretation, Geophys. J. R. astr. Soc., 51, 275-312.

Cisternas, A., Betancourt, O., and Leiva, A., 1973. Body waves in a "real earth" - Part I, Bull. Seism. Soc. Am., 63, 145-156.

Cleary, J. R., 1967. Analysis of the amplitudes of short-period P waves recorded by long range seismic measurement stations in the distance range 30° to 102°, J. Geophys. Res., 72, 4705-4712.

Cleary, J. R., 1974. The D" region, Phys. Earth Planet. Int., 9, 13-27.

Cormier, V. F. and Richards, P. G., 1977. Full wave theory applied to a discontinuous velocity increase: The inner core boundary. J. Geophys. Res., 43, 3-31.

Doornbos, D. J., 1978. On seismic wave scattering by a rough core-mantle boundary, Geophys. J. R. astr. Soc., 53, 643-662.

Doornbos, D. J. and Husebye, E. S., 1972. Array analysis of PKP phases and their precursors, Phys. Earth Planet. Int., 5, 387-399.

Doornbos, D. J. and Mondt, J. C., 1979a. Attenuation of P and S waves diffracted around the core, Geophys. J. R. astr. Soc., 57, 353-379.

Doornbos, D. J. and Mondt, J. C., 1979b. P and S waves diffracted around the core and the velocity structure at the base of the mantle, Geophys. J. R. astr. Soc., 57, 381-395.

Dziewonski, A. M., Hales, A. L, and Lapwood, E. R., 1975. Parametrically simple earth models consistent with geophysical data, Phys. Earth Planet. Int., 10, 12-48.

Elasser, W. M., Olson, P., and Marsh, B. D., 1979. The depth of mantle convection, J. Geophys. Res., 84, 147-155.

Gilbert, F., and Dziewonski, A., 1975. An application of normal mode theory to the retrieval of structural parameters and source mechanism from seismic spectra, Phil. trans. R. Soc. London, Ser. A, 278, 187- .

Gilbert, F. and Helmberger, D. V., 1972. Generalized ray theory for a layered sphere, Geophys. J. R. astr. Soc., 27, 57-80.

Gutenberg, B., 1960. The shadow of the earth's core, J. Geophys. Res., 65, 1013-1020.

Haddon, R. A. W., 1978. Vital P wave scattering in the earth

(abstract), XII IUGG symposium on mathematical geophysics, Caracas, Venezuela.

Haddon, R. A. W. and Cleary, J. R., 1973. A note on the interpretation of precursors to PKP. Phys. Earth Planet. Int., 7, 495-497.

Helmberger, D. V., 1968. The crust-mantle transition in the Bering Sea, Bull. Seism. Soc. Am., 58, 179-214.

Helmberger, D. V. and Wiggins, R. A., 1971. Upper mantle structure of the midwestern United States, J. Geophys. Res., 76, 3229-3245.

Husebye, E. S., King, D. W., and Haddon, R. A. W., 1976. Precursors to PKIKP and seismic wave scattering near the mantle-core boundary. J. Geophys. Res., 81, 1870-1882.

Jeanloz, R., and Richter, F. M., 1979. Convection, composition, and the thermal state of the lower mantle, J. Geophys. Res., 84, 5497-5504.

Johnson, L. R., 1969. Array measurements of P velocities in the lower mantle, Bull. Seism. Soc. Am., 59, 973-1008.

Jones, G. M., 1977, Thermal interaction of the core and the mantle and long term behavior of the geomagnetic field, J. Geophys. Res., 82, 1703-1709.

Kanamori, H., 1967. Spectrum of P and PcP in relation to the mantle-core boundary and attentuation in the mantle, J. Geophys. Res., 72, 559-571.

Kennett, B. L. N., 1974. Reflections, rays and reverberations, Bull.

Seism. Soc. Am., 65, 1643-1651.

King, D. W., Haddon, R. A. W., and Cleary, J. R., 1973. Evidence for seismic wave scattering in the D'' layer, Earth Planet. Sci. Lett., 20, 353-356.

Krebes, E. S. and Hron, F., 1980. Ray-synthetic seismograms for SH waves in anelastic media, Bull. Seis. Soc. Am., 70, 29-46.

Langston, C. A. and Helmberger, D. V., 1975. A procedure for modelling shallow dislocation sources. Geophys. J. R. astr. Soc., 42, 117-130.

Mitchell, B. J. and Helmberger, D. V., 1973. Shear velocities at the base of the mantle from observations of S and ScS, J. Geophys. Res., 78 6009-6020.

Muller, G., 1970. Exact ray theory and its applications to the reflection of elastic waves from vertically inhomogeneous media, Geophys. J. R. astr. Soc., 21, 261-283.

Muller, G., 1971. Approximate treatment of elastic body waves in media with spherical symmetry, Geophys. J. R. astr. Soc., 23, 435-449.

Phinney, R. A. and Alexander, S. S., 1969. The effect of a velocity gradient at the base of the mantle on diffracted P waves in the shadow, J. Geophys. Res., 74, 4967-4971.

Phinney, R. A. and Cathles, L. M., 1969. Diffraction of P by the core: A study of long-period amplitudes near the edge of the shadow, J. Geophys. Res., 74, 1556-1574.

Rial, J. A. and Cormier, V. F., 1980. Seismic waves at the

epicenter's antipode, J. Geophys. Res., 85, 2661-2668.

Richards, P. G., 1971. Waves in stratified media, Geophysics, 36, 798-809.

Richards, P. G., 1973. Calculation of body waves, for caustics and tunneling in core phases, Geophys. J. R. astr. Soc., 35, 243-264.

Richards, P. G., 1976. On the adequacy of plane-wave reflection/transmission coefficients in the analysis of seismic body waves, Bull. Seism. Soc. Am., 66, 701-717.

Ruff, L. J. and Anderson, D. L., 1980. Core formation, evolution, and convection: A geophysical model, Phys. Earth Planet. Int., 21, 181-201.

Ruff, L. J., and Butler, R., 1980. Teleseismic short-period amplitudes: path effects, in preparation.

Sacks, I. S., 1966. Diffraction wave studies of the earth's core, J. Geophys. Res., 71, 1173-1181.

Sharpe, H. N., and Peltier, W. R., 1979. A thermal history model for the Earth with parameterized convection, Geophys. J. R. astr. Soc., 54, 171-204.

Teng, T. L., 1966, Body-wave and earthquake source studies, Ph.D. thesis, California Institute of Technology, Pasadena, California.

Teng, T. L., 1968. Attenuation of body waves and the Q structure of the mantle, J. Geophys. Res., 73, 2195-2208.

Turcotte, D. L., and Oxburgh, E. R., 1967. Finite amplitude convective cells and continental drift, J. Fluid Mech., 28,

24-42.

Woodhouse, J. H., 1978. Asymptotic results for elastodynamic propagator matrices in plane-stratified and spherically-stratified earth models, Geophys. J. R. astr. Soc., 54, 263-280.

Appendix A for Part II

The purpose of this appendix is to outline the generalized ray method and to then show how anelasticity can be incorporated. There is a comprehensive literature concerning the details of both the generalized ray and full wave methods. The generalized ray method is developed with tarious applications in: Helmberger (1968), Chapman (1974a), Langston and Helmberger (1975), and Bwrdick and Helmberger (1978). The full wave method in the seismological context is developed in: Richards (1971), Richards (1973), Chapman (1974b), and Richards (1976). Examples of applications are: Choy (1977), Cormier and Richards (1977), and Rial and Cormier (1980). The earth flattening transformation, diffracted waves, and ray expansions are discussed in, for example: Muller (1971), Chapman (1973), Chapman and Phinney (1972), Muller (1970), Cisternas et al., (1973), Kennett (1974), and Woodhouse (1978).

Generalized ray method

Starting with spherical geometry and the physical properties varying only with radius r, a sequence of transformations applied to the elasto-dynamic equations results in the stresses and displacements having the dependence, $Y(r,p,s)$, where p is the ray parameter and the time dependence is Laplace (s) or Fourier (ω), transformed. For convenience in matching boundary conditions at specified values of r, the displacements and tractions are organized into a stress-displacement

vector, $\bar{V}_s$. The P-SV and SH wave systems decouple, and the P-SV stress-displacement vector has four components and is given by:

$$\frac{d\bar{V}_s}{dr} = M_s\bar{V}_s + \bar{w}_s\ (r_w),\qquad\qquad (A.1)$$

where the matrix M_s contains the varying physical properties, $\bar{w}_s$ is the source description at radius r_w and the subscript s refers to the spherical geometry. Synthetic body wave methods start with the above system.

To put equation (A.1) in a form suitable for the Cagnaird-de Hoop transform inversion: an earth flattening transformation is applied, the earth structure is converted from a continuously varying structure to a stack of thin homogeneous layers, then a ray expansion is used to isolate the arrivals of interest. The earth flattening transformation used for the P-SV system results in a high-frequency approximation to equation (A.1) (Chapman, 1973). By representing the structure as a stack of homogeneous layers, M is constant within each layer, and for layer index k with $Z_j \leqslant Z \leqslant Z_k$ ($d_k = Z_k - Z_j$), the flattened earth form of equation (A.1) is:

$$\frac{d\overline{V}_k}{dZ} = M_k \overline{V}_k \; . \tag{A.2}$$

If the source depth Z_w is within layer k, then the source term is added to equation (A.2). The solution to a system of equations such as equation A.2 is well-known. Let D_k be the diagonal form of M_k, i.e., $M_k = N_k D_k N_k^{-1}$. Then, for a particular layer, the solution is:

$$\overline{V}_k(Z_j) = P_k(Z_j, Z_k) \overline{V}_k(Z_k) \quad ,$$

where the propagator matrix is,

$$P_k(Z_j, Z_k) = N_k \exp \left[D_k(Z_j - Z_k) \right] N_k^{-1} \quad .$$

Recalling that the stress-displacement vector is continuous across the boundaries, the solution for the layered structure can be written as a contraction of the propagator matrices,

$$\overline{V}(Z_o) = P_1(Z_o, Z_1) \; P_2(Z_1, Z_2) \; . \; . \; . \; P_k(Z_j, Z_k) \overline{V}_k(Z_k) , \tag{A.3}$$

where each propagator matrix propagates the stress-displacement vector across the layer. A ray of expansion of equation (A.3) can be developed by noticing that the eigenvalues of D correspond to the four wave types: upgoing and downgoing P and SV waves. The displacement at Z_k due to the direct P wave can be isolated by considering only the P wave segment in each layer, e.g., $\exp(D_1 d_1) \rightarrow \exp(s \eta_{p1} d_1)$, and extracting the transmission/reflection coefficients for p waves from the contraction of $N_k^{-1} N_j$. Therefore, the vertical component of displacement at $Z = Z_o$ due to the P wave ray that propagated down from the source at $Z = Z_o$, is reflected at $Z = Z_k$, and propagates back up to Z_o is,

$$v_k(Z_o, p, s) = g(p, s) T_k R_k \exp(-2s \sum_{j=1}^{k} \eta_{pj} d_j) , \qquad (A.4)$$

where T_k is the product of the downgoing and upgoing P-wave transmission coefficients, R_k is the P-wave reflection coefficient at $Z = Z_k$, and the factor g contains the source description and displacement receiver factor. The total response at Z_o due to the primary P-waves is obtained by summing the reflected generalized rays over the relevant depth range,

$$v_p(Z_o, p, s) = \sum_{k=n}^{m} v_k(Z_o, p, s) \ . \tag{A.5}$$

The inverse transform for $p \rightarrow x$ in the asymptotic form is,

$$v_p(Z_o, x, s) = \text{Im} \int_0^{i\infty} v_p(Z_o, p, s) \exp(-spx) \ dp \ . \tag{A.6}$$

The Cagnaird-de Hoop inversion technique manipulates the above integral so that it appears to be the LaPlace transform

$$v(Z, x, s) = \int_0^{\infty} v(Z, x, t) \exp(-st) \ dt \ . \tag{A.7}$$

$v_p(Z_o, x, t)$ is then extracted directly from the integrand of equation (A.6). This requires placing the s dependence of $v_p(Z_o, p, s)$ into the exponential, $\exp(-spx)$, thus identifying t in (A.7) as $px + 2\Sigma \eta_{pj} d_j$. Any other s dependence can be absorbed into the source description, such that $g(p,s) \rightarrow g(p)$. Also, the variable of integration is changed, $dp = (dp/dt)dt$, and the p contour must be deformed so that $\text{Im}[t(p)] = 0$, as time must be real. Applying these changes,

$$v_k(Z_o, x, t) = \text{Im}[g(p)T_k R_k \, dp/dt] \; .$$

(A.8)

The turning ray in the inhomogeneous spherical model corresponds to the generalized ray which reflects at the critical angle, i.e., the minimum time arrival. Complicated velocity structures and multiple turning points cause no difficulties in calculating the individual generalized rays. However, when waveguides are present, one must insure that the number of rays used is adequate, or if multiply reflected P-waves should be added to the ray sum. Diffraction by the core occurs when the distance, x, is beyond the distance of the critically reflected ray just above the CMB. Consequently, the diffracted response is composed of head waves.

Anelastic generalized ray method

The effects of anelasticity can be introduced into equation (A.6) by allowing the velocity to have a small imaginary part. That is, the P-wave velocity can be written as $\alpha^* = \alpha + ia$. This perturbation approach is valid for $Q^{-1} \ll 1$, where Q is the seismic quality factor. For small Q^{-1}, most of the attenuation effect is due to the imaginary part of the exponential argument in equation A.6. The vertical slowness η_p is given by $\eta_p = (1/\alpha^2 - p^2)^{1/2}$. Letting $\epsilon = a/\alpha$, the complex vertical slowness is,

$$\eta_p{}^* \cong \eta_p \left(1 + \frac{i2\epsilon}{\alpha^2\eta_p{}^2}\right)^{1/2} \cong \eta_p - i\frac{\epsilon}{\alpha^2\eta_p} \quad , \tag{A.9}$$

where η is the real vertical slowness. Substituting η^* for η in equation (A.6), we obtain an additional term in the exponent, i.e.,

$$\exp\left(-is\, 2\sum d_j\epsilon_j/\alpha^2{}_j\eta_{pj}\right) \quad .$$

To use the Cagnaird-de Hoop method as before, we want to take the attentuation factor out of the p integral. This can be accomplished by noticing that $d_j/\alpha_j{}^2\eta_{pj}$ is the one-way travel time through layer j. Therefore, with $\epsilon_j = Q_j{}^{-1}/2$, the attenuation factor is $\exp[-i(s/2)\Sigma t^*{}_j]$, where $t^*{}_j = $ (travel time) $_j/Q_j$. Thus, with no p dependence, the attenuation factor can be taken out of the p integral. Although the attenuation factor depends upon s, the Cagnaird-de Hoop inversion can still be used, later convolving the inverse Laplace transform of the attenuation factor with the result, i.e., $y(s)=u(s)v(s)$, then $Y(t)=U(t)*V(t)$. Let

$$t^*_k = \sum_{j=1}^{k} t^*_j$$

then including the attenuation factor in equation (A.6),

$$v_p(Z_o, x, s) = \sum_{k=n}^{m} \exp\left(-i\,\frac{s}{2}\,t^*_k\right) \text{Im} \int_0^{i\infty} gT_k R_k \exp\left[-s\left(px + 2\sum_{j=1}^{k} \eta_{pj} d_j\right)\right] dp.$$

(A.10)

Usually it is assumed that t^*_k is nearly the same for the reflection depth range of $k = n$ to $k = m$, in which case the attenuation factor is taken outside the sum and the attenuation operator with a single value of t^* is convolved with the total elastic response, v_p. However, if t^*_k varies significantly over the depth range of interest, it is necessary to convolve a different t^* operator with each individual generalized ray, v_k. The LMQ1 model requires the latter treatment.

It is interesting to note the approximations implied when assuming that all of the attenuation effect occurs in the exponential factor, then taking this factor out of the p integral. The attenuation effect on the reflection coefficient is ignored. Anelasticity affects the reflection coefficient, but it is important only for critical incidence with a very low Q, [10-20 (Borchedt, 1977; Krebes and Hron, 1980)].

Taking the attenuation factor outside the p integral assumes that t^*_k remains real throughout the response. Thus, t^*_k is the proper value for the geometrically reflected ray off interface k. Clearly, p acquires an imaginary part as the de Hoop contour leaves the real axis, and t^*_k then has an imaginary part. This detail is not important for

the pre-critical reflection. When the reflection from interface k is beyond the critical angle, then p has a small imaginary part corresponding to the head wave contribution. Our approximation for t^*_k ignores this aspect, both the modification to the attenuation factor and the second order phase term that appears. If the Q is the same on both sides of the interface, then this approximation causes a very small error. To crudely estimate the maximum error introduced in the calculations for LMQ1, let x_H be the distance beyond the distance of the critical reflection from a particular interface (α_1, Q_1 above, and α_2, Q_2 below). Also, let dt^*_H be the correction that should be added to the reflected t^*. Then, for a small velocity contrast and thin layers,

$$dt^*_H \sim \left[(1/\alpha_2 Q_2) - (1/\alpha_1 Q_1) \right] \; x_H \, .$$

Thus, for an interface with a factor of 5 contrast in Q (as in LMQ1), at $\Delta = 105$ deg., $dt^*_H \sim 0.3$ sec. As the Q is constant within the low-Q zone of LMQ1, dt^*_H is negligible at the other interfaces in the low-Q zone. Therefore, we do not expect the total differential t^* for the head waves to accumulate significantly at $\Delta \sim 105$ deg., and the amplitude error at $\Delta \sim 105$ deg. will be substantially less than 50%.