Correlation-Based Phase Pick Correction and Similar Earthquake Family Identification in Large Seismic Waveform Catalogues

by
Charlotte A. Rowe

Submitted in Partial Fulfillment of
the Requirements for the Degree of
Doctor of Philosophy in Geophysics
Department of Earth and Environmental Science
New Mexico Institute of Mining and Technology
Socorro, New Mexico

October, 2000

ABSTRACT

This thesis describes the development of an automatic, adaptive algorithm which improves the consistency of picking arrival times of seismic phases in digital seismic waveform catalogues and calculates quantitative errors of these picks. The resulting improved pick conssistency reduces the random component of earthquake hypocenter location and permits better definition of interpretable seismogenic structures; the improved picks and their associated standard deviations may also be useful in other applications which depend upon these parametric data.

Pick adjustments are obtained by comparing station-common waveforms that have preliminary picks, and estimating relative lags between these picks. Relative lag estimation between pairs of traces proceeds in two steps: an integer-sample correction is made, followed by a sub-sample refinement. For stations with multiple components, we first use polarization filtering to improve signal-to-noise levels prior to waveform comparison. Automatic, adaptive, cross-coherency filtering is then applied to the waveform pair. A suite of cross-correlation functions is calculated from narrow-band-filtered representations of the waveforms to estimate integer correlation lags and associated standard deviations.

The technique uses a cross-spectral phase slope method for subsample estimates. Multitaper spectral estimation provides the best estimate of the cross-spectrum, as well as statistically and dimensionally meaningful estimates of the subsample standard deviation. Summing the integer and sub-sample lag estimates provides the total pick adjustment for each waveform pair. The associated pick standard deviation is estimated as the quadrature

sum of the coarse and fine standard deviations. The resulting system of inter-event constraints is solved using an iterative Polak-Ribiere conjugate gradient minimization of the L1-norm residual, refined by conservatively rejecting outliers based on interim solutions and successively resolving the system. After a satisfactory solution is found we obtain final pick adjustments for the reduced system and calculate 1-sigma error bars on this solution via Monte Carlo propagation of Gaussian data errors.

A hierarchical, dendrogram-based clustering method is employed to separate large catalogues into similar event families, each of which may be solved independently for intra-cluster relative pick adjustments. A technique of stacking and stack correlation adjusts inter-cluster picks for consistency within the full catalogue.

The algorithm is applied to two different waveform catalogues. Manually repicked waveform data from a subset of microearthquakes at the Soultz-sous-Forets geothermal reservoir, Alsace, France, are used to test the algorithm against careful human repicking, then the technique is applied to a ~7,000-event dataset from the reservoir. Relocated events exhibit significantly more structure than was seen in the preliminary hypocentral cloud, and delineate numerous intersecting joint features in the reservoir. In a second application, some 4,000 volcanic earthquakes associated with the 1989-1990 eruption of Redoubt Volcano, Alaska are relocated. These relocations show significant improvement in location consistency for volcanic long period events arising from a stationary volumetric source, and show increased clustering of volcanotectonic events into discrete seismically active clusters as well as fractures or joints at depth. Routine use of the technique may enhance observatory operations and improve monitoring capabilities at dangerous volcanoes.

ACKNOWLEDGEMENTS

This work owes its completion to my advisor, Rick Aster, who provided scientific and strategic guidance, as well as extraordinary opportunities to meet with, talk to, and learn from many people, both at home and abroad. Thanks also to Brian Borchers, for invaluable guidance in debugging, testing and optimizing these codes. My other committee members, John Schlue and Harold Tobin, were helpful in providing critiques of writing and analysis as well as discussions and advice along the way; you all have been a great committee. Thanks also to Philip Kyle for the time served as a part of the committee.

This work was supported through grants from Niitsuma Laboratories, Tohoku University, Sendai, Japan, Sandia National Laboratories and the National Science Foundation grant #OPP-9419267. Scott Phillips and Michael Fehler of Los Alamos National Laboratory and Christopher Young of Sandia National Laboratories provided helpful insights and discussions. I am grateful to the AGU, NMT Graduate Student Association, Fred Philips and Van Romero for travel support to international meetings that proved pivotal in the all-important quest for securing post-doctoral positions and additional collaborative projects.

Nobody gets through a Ph.D. alone, despite grand illusions of independence and self-sufficiency. Moral support, cheerleading, wine, dinner, skiing, chocolate, pet care and patience were graciously dished out and/or shared by my parents, neighbors, and many friends from Socorro to Alaska to the far-flung corners of the planet. Thank you all.

TABLE OF CONTENTS

LIST OF FIGURES

Page

Part A: Relocation of Seismicity Associated with the 1989-1990 Eruption of Redoubt Volcano, Alaska: Part I. An Automatic, Adaptive Algorithm for Refining Phase Picks in Large Seismic Data Sets

Part B: Relocation of Seismicity Associated with the 1989-1990 Eruption of Redoubt Volcano, Alaska: Part II. Application to Redoubt Eruption Seismicity

Part C: Relocation of Induced Microearthquakes at the Soultz Geothermal Reservoir Using Automated, High-Precision Re-picking

LIST OF TABLES

This dissertation is accepted on behalf of the faculty of the Institute by the following committee

Adviser

23 October, 2000
Date

I release this document to New Mexico Institute of Mining and Technology.

Student's Signature

23 October 2000
Date

INTRODUCTION

The primary aim of this project has been the development of an automatic computer algorithm which can be applied to large, digital seismic waveform catalogues to correct for phase pick inconsistencies among similar earthquakes. Software already exists to address this problem, but the approaches adopted by other researchers did not satisfy such priorities as: computational and operational efficiency, and easy adaptability for general use and portability. Standard signal processing techniques such as cross-correlation and cross-spectral correlation methods are modified to incorporate a new method of automatic, adaptive filtering based on cross-coherency. Automatic polarization filtering is incorporated for any multicomponent data sets. The method of spectral estimation using multiple eigentapers is adapted to help determine robust cross-spectral measurements as well as dimensionally meaningful uncertainty estimates. The method of conjugate gradient solution for the system of inter-event lag equations is implemented for L1-norm residual minimization, with modifications to avoid discontinuities in the derivative, which become problematic with near-zero misfit values. A clustering algorithm is employed to provide flexible similarity classification, which depends upon the characteristics of an individual dataset.

The algorithm was developed incrementally, through comparison of the computer repicks against careful human repicking of a small subset of events from the Soultz geothermal reservoir in France. Once automatic relocations provided a satisfactory compari-

son to the manual relocations, a larger dataset of ~7,000 Soultz reservoir events was addressed. The method was then applied to seismicity recorded during the 1989-1990 eruption of Redoubt Volcano (Alaska).

This dissertation is divided into three sections, each of which addresses a different aspect of the project. In addition, a user's guide for the algorithm is included.

The first two sections of the dissertation, A and B, are being submitted to the Journal of Geophysical Research as companion papers. These are:

Relocation of Seismicity Associated with the 1989-1990 Eruption of Redoubt Volcano, Alaska:
Part I. An Automatic, Adaptive Algorithm for Refining Phase Picks in Large Seismic Data Sets
Part II. Application to Redoubt Eruption Seismicity

The first of these papers outlines the repicking algorithm. Authors on the submitted manuscripts are C.A. Rowe, R.C. Aster, B. Borchers and C.J. Young. My co-authors contributed significantly to the development of the computer algorithm. R. Aster crafted an early version of the correlator and wrote some Matlab test implementations for ideas which were later incorporated into the C programs, both for correlator and conjugate gradient solver. B. Borchers provided critical guidance for program optimization and debugging, as well as suggestions for better tools such as the FFTW routines; he performed the final debugging of the conjugate gradient solver. C. Young provided the Matlab routines on which the C clustering programs were based. This section has an appendix, which is a user guide to the (current version of the) software.

The second of the companion papers is an application of the repicking algorithm to seismicity associated with the 1989-1990 eruption of Redoubt Volcano, Alaska. In this paper, the method not only provides relocations of different types of volcanogenic seismic

events, but demonstrates some unexpected insights into the nature of the volcanic seismic sources as well as possible seismic network diagnostic applications.

The third part of the dissertation,

Relocation of Induced Microearthquakes at the Soultz Geothermal Reservoir Using Automated, High-Precision Re-picking,

has been submitted to the journal *Pure and Applied Geophysics*, as an invited contribution to a special volume on induced microearthquakes (edited by Cezar Trifu). Authors on this paper are C. Rowe, R. Aster, W.S. Phillips, R. Jones, B. Borchers and M. Fehler. My co-authors provided the data, support and very helpful discussions regarding the nature of these data. R. Jones provided unpublished or otherwise inaccessible results from related studies (such as rupture dimensions and velocity model development) upon which I base some of my own analysis. Without the collaboration of R. Aster and B. Borchers, the repicking tools used in the Soultz application would not have been developed.

Part A Relocation of Seismicity Associated with the 1989-1990 Eruption of Redoubt Volcano, Alaska: Part I. An Automatic, Adaptive Algorithm for Refining Phase Picks in Large Seismic Data Sets:

ABSTRACT

Significant advances have been made recently in the arena of precise earthquake relocation by applying signal processing methods to the growing worldwide catalogues of digital waveform data. Techniques are becoming increasingly efficient, with improved ability to address even the largest digital waveform catalogues, yet some significant degree of analyst intervention generally is still required. Furthermore, those techniques addressing the largest data sets make *a priori* assumptions about waveform similarity as a function of inter-event distances, depending upon preliminary hypocenter locations to establish weighting parameters or to maintain absolute location centroids for identified multiplet groups.

We have developed an adaptive, automatic, correlation- and clustering-based method for greatly reducing the degree of picking inconsistency in large, digital seismic catalogues and for quantifying similarity within clusters of disparate yet spatially coincident waveform families. Innovations in the technique include the use of eigenspectral methods for cross-spectral phase estimation and to provide statistically and dimensionally meaningful pick lag error estimates, adaptive coherency-based filtering, and a hierarchical clustering and stacking method for adjusting inter-cluster centroids without compromising tight intra-cluster relative lag estimates. To solve the systems of lags we employ an opti-

mized conjugate gradient technique that minimizes an L_1-norm misfit. Our repicking technique provides robust similarity classification and improved phase arrival estimation for situations with sparse network coverage and heterogeneous source regions.

INTRODUCTION

Earthquake location and many other travel-time-based seismological applications have historically depended critically upon the ability of human analysts to obtain good arrival times of body waves. Standard network operations generally involve the manual measuring of P-wave and S-wave arrivals or, more recently, computer identification of these phases using software autopickers. The most common human or computer picking approach is done one event at a time, with records from several or all recording stations. These methods, although well suited for the near-real-time processing demands of network operation, does not produce consistent phase arrival times (picks), because path effects, signal to noise conditions and source radiation pattern differences within the network of receivers may be large. Human (or automatic computer) picks obtained from the resulting heterogeneous suite of waveforms may thus be highly inconsistent for even very similar events. These inconsistent picks are then used to calculate the hypocenter, and the event is archived. Seldom are any but the most egregious picking errors noted and corrected prior to moving on to the next earthquake; hence, picking inconsistencies between even similar events remain unresolved.

Such routine network operations have produced very large sets of hypocenter locations (e.g., in excess of four million events with over a hundred million associated picks from 1981 - 1999 for southern California) with location error estimates of generally a few to a few tens of kilometers for regional-scale networks with good azimuthal coverage.

Standard catalogue locations have served to document general seismicity levels and the gross geometry of comparably-scaled seismogenic features but, within the diffuse scatter of locations which result from picking inconsistencies, fine details remain unresolved. Significant improvement in the precision of hypocenter location, and resulting delineation of the details of seismic source regions, has sometimes been achieved through careful, painstaking visual cross-correlation and repicking of phases for preliminarily located events (e.g., Phillips et al., 1997; Fehler et al., 2000; Phillips, 2000), but this is generally a time- and cost-prohibitive undertaking which will necessarily be limited to small, focused subsets of the larger catalogues.

Quantitative, waveform correlation-based phase repicking and/or relative doublet and multiplet relocations have produced some impressive resolution of seismogenic structures within families of similar events. Fremont and Malone (1987) used relative relocations based on cross-correlation lags of multiplets at Mount St. Helens to delineate source regions on the order of a few tens of meters. Deichmann and Garcia-Fernandez (1992) used cross-correlation methods to identify relative arrival time differences among similar Alpine events for precise relative location. Got et al. (1994) relocated seismicity at Kilauea volcano, Hawaii, using multiplets chosen based on cross-spectral coherency. Slunga et al. (1995) used relative arrival times calculated from Fourier-interpolated cross-correlation functions to determine precise relative locations and improved absolute locations in clusters of similar microearthquakes in Iceland by incorporating the cross-correlation lags into a modified Joint Hypocentral Determination (JHD) application. Gillard et al. (1996) used cross-correlation methods and multiplet analysis at Kilauea, Hawaii, to reduce quasi-linear "cigars" of microearthquake foci into precise relative

relocations delineating pencil-thin lines of seismicity. These trends were interpreted to correspond to growing stress concentration in the upper East Rift arising in response to seaward migration of the south flank of the volcano.

Dodge et al. (1995) developed a semi-automatic, computer-based correlation approach which calculates individual pick time inconsistencies for event pairs, and uses the weighted lag constraints to adjust for consistent picks. This approach is a significant improvement over earlier efforts which rely upon a master event method, in that master-event bias is reduced, highly similar multiplets are not required, and dissimilar event pairs will have limited influence over the pick adjustments. Shearer (1997) demonstrated significant improvement in delineating the seismogenic features associated with the Whittier Narrows aftershock sequence in California by also invoking a pick lag estimation method after first segregating events into groups meeting minimum cross-correlation criteria (e.g., Aster and Scott, 1990). Adjusting the phase picks, rather than simply relocating hypocenters based on calculated traveltime differentials, also provides an advantage over some other correlation-based event relocation methods, in that the corrected picks may not only be used for event relocation, but may be used in other applications such as seismic tomography (e.g., Kissling, 1988).

The above methods, although successively more quantitative and efficient, still rely to some significant degree on user interaction and have been so far applied to specific studies of catalogue subsets chosen either through spatial restrictions or genetic assumptions (e.g., a limited box of data, a specific aftershock sequence), or they operate by preliminary exclusion of individual events failing to meet very high (~0.9) cross-correlation criteria. A certain *a priori* selection to ensure correlatability has therefore

been invoked. We describe a method which combines the most desirable features of the techniques already available with additional adaptability and portability, in an automatic package which can be implemented for large catalogs such that a wide variety of applications may be addressed with little need for time-consuming customization.

We first outline the technique with a discussion of the signal processing tools incorporated in our algorithm and an explanation of how we apply them to the data. The problem of catalogue segregation and identification of similar event families will then be discussed. We then outline the final calculation of lags and standard errors from the inter-event constraints. This part of the dissertation is the first of a pair of companion manuscripts being submitted to the Journal of Geophysical Research (Rowe et al., 2000a, 2000b).

TECHNIQUE

Preliminary Data Organization

Events are first arranged into individual event directories, each of which contains all available traces as individual time series files. We currently use SAC data file format (Goldstein et al., 1998). Prior to processing, waveforms have preliminary P and/or S picks produced by an analyst or autopicker. These parameters are read from the trace headers, along with other potentially useful parametric data such as preliminary hypocenter coordinates.

As with other waveform correlation methods for precise earthquake relocation, we simultaneously analyze traces from many events on a station by station basis (e.g. Dodge et al., 1995; Shearer, 1997). This data regrouping allows one to improve the consistency of pick times among events by exploiting waveform similarity. The tightest constraints can

be expected to be found among the most similar earthquake waveforms, which arise from events having the most consistent radiation patterns, source time functions, and propagation paths. This approach is somewhat similar to the statics correction problem in reflection seismology (e.g., Marsden, 1993; Sheriff and Geldart, 1995), although in reflection applications the spatial sampling is sufficient to limit path effect variations between nearby receivers so that processing may still be done on an event-by-event basis.

Relative Lag Estimation

Relative lag estimation between pairs of traces proceeds in two steps; a coarse discrete correlation step, which provides an estimate of lag to the nearest time sample, and a fine correlation step which provides a refinement to the sub-sample level. Among all events recorded at a given station, we compare each event pair for each phase (P, S) to measure waveform similarity and to estimate lag. A user-specified M-sample time window, including a fractional pre-pick offset, is established about the preliminary picks for each pair of events to be compared. This windowlength is chosen based upon sample rate and overall frequency content of the targeted phase. Generally speaking, two cycles of the dominant waveform is acceptable; however, the choice of correlation window length varies depending upon the intended use of the correlation results. In our application, we select a window which includes several cycles for the first cross-correlation, the results of which will determine similarity clusters. Following cluster separation, intra-cluster cross-correlations are performed using a suite of correlation windowlengths. This range is chosen such that the minimum window length includes one to two cycles of the highest frequency component that may be consistently identified among a suite of sample waveforms; the longest window is chosen based upon a conservative estimate of likely maximum pick error.

These adaptive windowlength ranges may be different for P-waves and S-waves, and will generally vary among stations. We select the optimum correlation result among the multiple window-length trials as that which exhibits the highest cross-correlation value combined with the lowest lag standard deviation (discussed below). In a companion paper (Rowe et al., 2000a), we explore the effect of short versus long correlation window choice on events having significantly different relative time lags between high-frequency and low-frequency constituents, and how this choice influences the resulting pick lag estimates.

For stations with multiple components (usually 3), we use polarization filtering to improve P and S signal-to-noise levels prior to waveform comparison. This provides the best function for subsequent correlation when the earthquake source is not favorably oriented for a particular component.

A mean covariance matrix (e.g., Aster et al., 1991) is calculated from the sum of the energy-normalized multicomponent signal for each event in the pair:

$$\mathbf{C} = \frac{1}{2}\left(\frac{\mathbf{x}_1^{\mathrm{T}}}{E_1} + \frac{\mathbf{x}_2^{\mathrm{T}}}{E_2}\right), \tag{1}$$

where $\mathbf{x}_j$ is the (usually) three-component matrix with columns which are individual component time series for event j, and

$$E_j = trace(\mathbf{x}_j^{\mathrm{T}} \cdot \mathbf{x}_j)^{\frac{1}{2}}. \tag{2}$$

The diagonalization of the positive definite $\mathbf{C}$ matrix gives the unit eigenvector $\hat{\mathbf{u}}_1$ characterizing the best data projection for mutually linearizing particle motion between the two traces. For two- or three-component seismograms, the eigenvalue/eigenvector decomposition of the signal covariance matrix may be calculated exactly; however, the increasing use of four-component (or more) sensors requires an iterative approach to the

signal decomposition. We have therefore made use of repeated Jacobi transformations (Press *et al.*, 1996) to find the principal components of the tensor. All subsequent analysis is performed on the projected (1-dimensional) data

$$\mathbf{x}'_j = \mathbf{x}_j \cdot \hat{\mu}_1 . \tag{3}$$

Each waveform pair is next transformed via Fast Fourier Transform (FFT) into the frequency domain

$$X_{jk} = \sum_{i=0}^{M-1} x'_{lj} e^{-ikl/N} . \tag{4}$$

The cross-spectrum,

$$s_k = X_{1k} X^*_{2k}, \tag{5}$$

(where * denotes complex conjugate)

and coherence,

$$c_k = \frac{\sum_{l=j-m}^{j+m} s_l}{\sum_{l=j-m}^{j+m} |s_l|}, \tag{6}$$

are calculated, where the coherence averaging width is $2m+1$ Rayleigh bins (incrementally decreasing near the zero and Nyquist frequencies). We choose m as a fraction of the windowlength, M, with a minimum of five Rayleigh bins.

Prior to cross-correlation, we use the coherence and signal power to uniformly prefilter the two seismograms under consideration to emphasize high frequencies that have high signal/noise, high coherency. The spectral weighting is

$$\gamma_k = (|X_{1k}||X_{2k}|)^{1/2} c_k . \tag{7}$$

This filtering thus downweights incoherent frequency bands while reducing removal of potentially useful signal (a risk in an *a priori* bandpass filter choice; Figure 1). Although the most coherent frequencies will most likely be found in the lower frequencies of the spectrum, the adaptive nature of this approach enhances the comparisons of highly

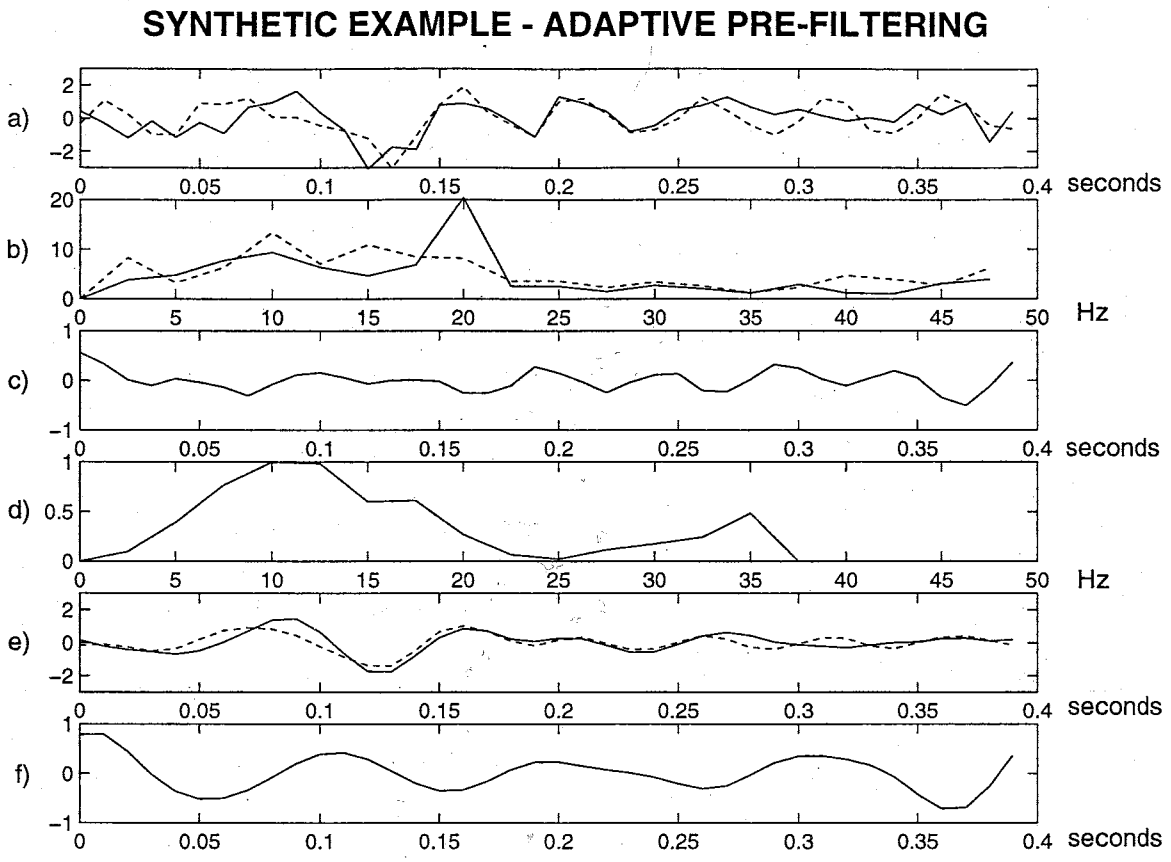


Figure 1. Example illustration of adaptive pre-filtering. a) Two synthetic seismograms with added Gaussian noise. b) Amplitude/frequency spectra of unfiltered seismograms from 1a. c) Cross-correlation function for unfiltered noisy seismograms in 1a; note that maximum cross-correlation coefficient is approximately 0.5. d) Cross-coherency for spectra shown in 1b, plotted as a function of frequency. Note that although most coherent energy resides below 20 Hz, cross-coherency has a small peak at 35 Hz. *A priori* lowpass filtering may reject this energy, which could be an important common constituent to include in the cross-correlation. e) Adaptively filtered seismograms from 1a; note significant reduction in the random noise constituent and overall similarity of resulting waveforms. f) Re-computed cross-correlation function for the waveform pair, showing maximum cross-correlation coefficient > 0.9. Initial cross-correlation provided a zero sample lag; the filtered trace pair yields a lag of 2 samples (from Aster and Rowe, 2000)

similar waveforms in a diverse catalogue by passing more of the coherent spectrum to each inter-event correlation function. At the same time, we can accommodate gross similarities among those event pairs whose higher frequency details may not correlate well.

The coherency-filtered signals, with spectra

$$Y_{jk} = X_{jk}\gamma_k, \tag{8}$$

are next cross-correlated (with zero padding to eliminate circular correlation wrap-around) in the frequency domain. The filtered cross-spectrum is transformed back into the time domain and the maximum of this cross-correlation function is the estimate for the coarse inter-event pick lag. Because initial pick inconsistency is sometimes an appreciable fraction of the window length, M, we symmetrically re-window the two traces (keeping track of the total window shift) until the final coarse lag between the re-windowed segments is 1, 0, or -1 sample. To estimate a standard deviation for the coarse (integer sample) correlation lags, l_α, we perform a set of narrow-band correlations (typically 8) and find the variance, where each term is weighted by the cross-spectral power in that band (Aster and Rowe, 2000). This coarse correlation standard deviation, σ_α, is used, with the coarse cross-correlation maximum, as a discriminator to determine which waveforms are sufficiently similar to merit being passed to the subsample cross-correlation step for further lag refinement.

In some instances, the desired level of relative lag resolution is less than the sample interval. For example, in an area where the P-wave velocity is approximately 5 km/s and where data are sampled at 100 sps, the integer sample arrival time resolution for even high-signal-to-noise data can be as poor as 0.005 s, which may introduce a worst-case location error of up to 25 m. The ability to consistently pick waveforms to subsample pre-

cision can dramatically improve resolution of small-scale features if sufficiently high quality data are available.

The cross-spectral method (e.g. Poupinet et al., 1984) provides a method for determining subsample lag adjustments by estimating a continuous function, the zero-intercept slope of the cross-spectral phase (5), where the sub-sample lag term estimate is

$$l_\beta = \frac{1}{2\pi}\frac{d\phi(f)}{df}, \tag{9}$$

and

$$\phi_k = \text{atan}\left(\frac{\text{imag}(s_k)}{\text{real}(s_k)}\right). \tag{10}$$

In many applications or this technique (e.g., Poupinet et al., 1984; Got et al., 1994), the relative weights of the phase values used in slope estimation have been calculated using a coherency-based measure. This provides useful relative weights to the phase points, but requires *ad-hoc* scaling to the standard deviations needed to estimate meaningful error statistics for the subsample lag. For re-estimation of phase picks and their subsequent inclusion in relocation or other applications, quantitative estimates of the pick standard deviations in time units are important.

Because accurate estimation of the cross-spectral phase slope is critical to obtaining high accuracy estimates of the relative lag and its standard deviation, we apply multitaper spectral estimation (Thompson, 1980). Spectral estimation on a sampled time series via FFT may be severely compromised when the length of the time window is shortened, as spectral leakage, which results from truncation of the windowing function, can bias the high-frequency roll-off of the spectral estimate (Park *et al.*, 1987) both in amplitude and phase. Yet for many seismological applications, including the cross-correlation-based repicking algorithm, it is desirable to use a fairly short time window to

isolate the limited and correlatable (direct phase arrival) portion of an intrinsically non-stationary signal; lengthening the window to reduce spectral leakage introduces a higher proportion of background noise, as well as scattering contributions from the coda, and degrades the direct phase arrival comparison.

A standard approach to reducing spectral leakage is to apply a taper to the truncated time series which smoothly downweights data points toward zero at the ends of the window. This provides good results in terms of reduced spectral leakage, but causes a severely elevated variance for the spectral estimate. The Hann taper, for instance, discards approximately 5/8 of the statistical information of the time series (Park et al., 1987). A further difficulty arises because of the intrinsic non-stationarity of the signal. Use of the Hann (or similar) taper tends to preferentially emphasize the spectral properties of that portion of the time series which falls in the central part of the window, while neglecting most of the information near the extrema, which may not be well represented by the center portion of the window.

One of the most successful approaches to solving these problems is the multitaper spectral estimation (Thomson, 1982), wherein discrete prolate spheroidal wave functions, which are eigenfunctions of the Dirichlet kernel, are employed. The eigenfunctions, denoted by $U_k(N, W; f)$, k=0,1...,N-1 are solutions to the equation:

$$\int_{-W}^{W} \frac{\sin N\pi(f-f')}{\sin \pi(f-f')} U_k(N, W; f')df' = \lambda_k(N, W) \cdot U_k(N, W; f), \qquad (11)$$

where W, $0<W<1/2$ is a bandwidth normally of the order $1/N$. The functions are ordered by their eigenvalues

$$1 > \lambda_0(N, W) > \lambda_1(N, W) > \ldots\ldots > \lambda_{N-1}(N, W) \qquad (12)$$

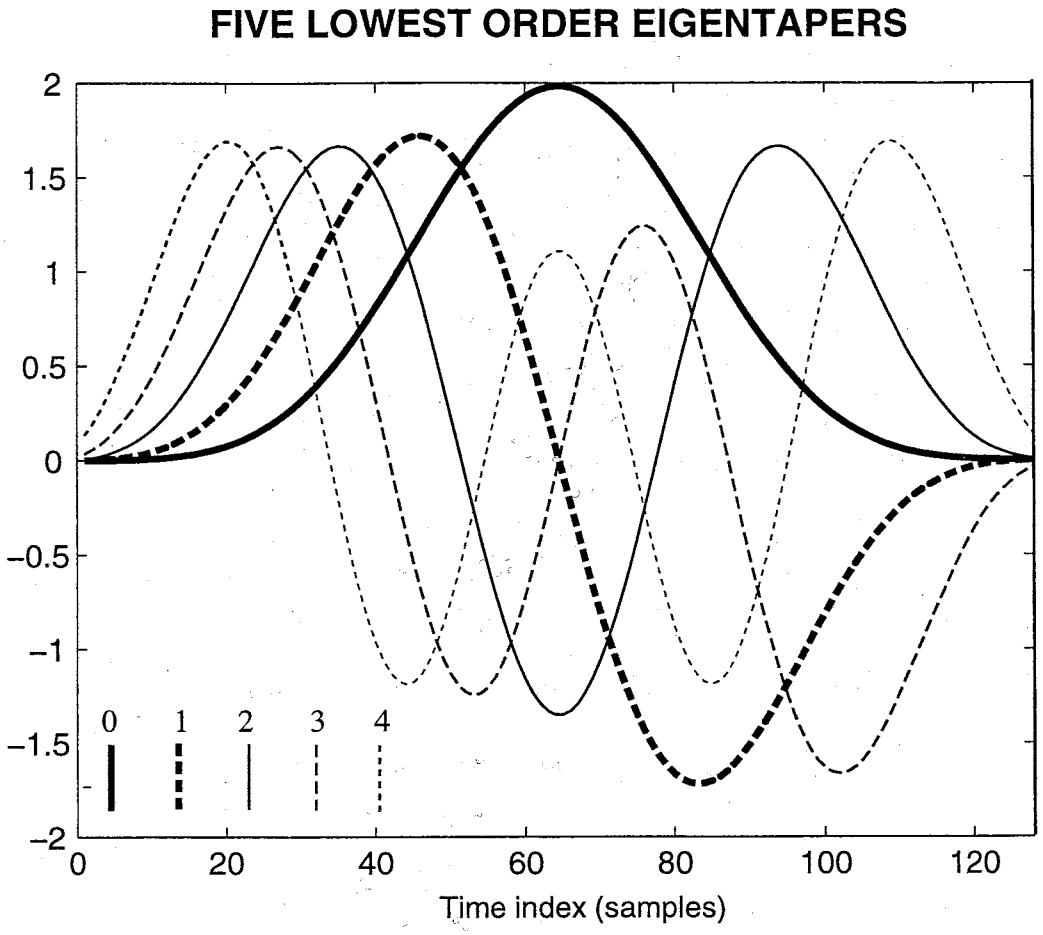


Figure 2. First five prolate spheroidal eigentapers for a time-bandwidth product of 4. Each taper recovers a different portion of the windowed seismogram; note that higher order tapers have increasingly steep initial slopes; hence, spectral leakage becomes greater with higher order tapers. The window is 128 samples long.

The first $2NW$ eigentapers have eigenvalues that are extremely close to 1. Of all functions which are the Fourier transform of an index-limited sequence, the discrete prolate spheroidal wave function has the greatest fractional energy concentration within the bandwidth of (-*W, W*) (Thomson, 1982). These eigenfunctions are orthogonal over the interval (-*W, W*) and are orthonormal over $(-1/2, 1/2)$. Their Fourier transforms provide the discrete prolate spheroidal sequences, also known as prolate eigentapers, with which we can window the time series prior to estimating its spectrum (Park *et al.*, 1987). We illustrate the five lowest-order multitaper functions for a 128-sample window in Figure 2.

The products of the time series with each of the eigentapers are mutually orthogonal, as are their Fourier transforms; hence, we obtain a linearly independent series of eigenspectra for the time series, which may be combined in a weighted sum to estimate its true spectrum. The orthogonality of the eigenspectra further permits us to calculate meaningful estimates of error statistics for the summed spectrum.

We pre-compute the tapering functions for a particular data length N and a specified time-bandwidth product, W. W is typically chosen to be 4, where $2W$ approximately specifies the resolution of the resulting spectral estimates in Rayleigh bins (Thompson, 1980; Park et al., 1987).

In our application, we calculate multitaper estimates of the cross-spectrum from two seismograms through conjugate multiplication (5) of corresponding multitaper spectra (Figure 3). Figure 3a shows two synthetic seismograms of length 32 samples. Each waveform is multiplied by the six lowest-order eigentaper functions, providing six linearly independent tapered realizations of each trace (Figure 3b). Each is then transformed to the

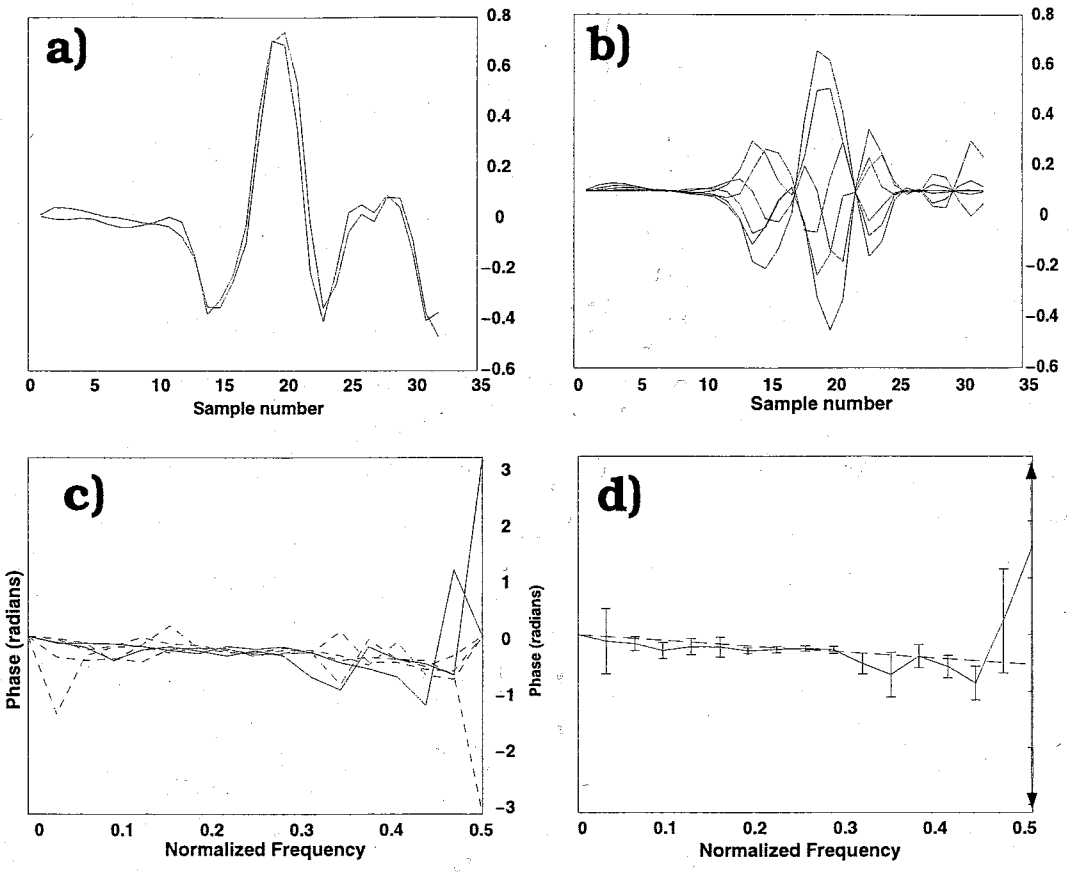


Figure 3. Example of estimation of cross-spectral phase slope and associated standard deviation. a) Two synthetic seismograms that have been aligned to the nearest sample with integer correlation. b) Six tapered representations of one of the synthetic traces; each function is the result of using one of the six lowest-order eigentapers to weight the time series. c) Six linearly independent estimates of the cross-spectral phase for the traces in Figure 3a. Each trace was windowed with the six lowest-order eigentapers (as in Figure 3b), then six corresponding cross-spectra were calculated. Solid lines represent the cross-spectra corresponding to the two lowest order eigentapers. The dashed functions are cross-spectral phases estimated from the third through sixth tapers. d) Mean cross-spectral phase function estimated using two lowest-order cross-spectra. Dashed line represents the phase slope estimate; vertical error bars show standard deviations for each Nyquist bin, estimated from all six cross-spectra. Final phase/frequency point with arrowed error bar has a standard deviation large enough to disqualify the point in the phase slope fitting.

frequency domain, and we compute six linearly independent cross-spectral phase estimates from the tapered functions (Figure 3c).

The lowest order eigentapers (especially 0 and 1) have very low spectral leakage outside of the specified spectral resolution bandwidth (k-W, k+W). Higher order tapers, however, have progressively worse spectral leakage characteristics which may unacceptably flatten the estimated phase slope, and hence underestimate the subsample correlation lag. This effect was seen to be especially significant for shorter window lengths (typically $N < 64$), where the maximum number of usable phase points is less than $K < (N/2 - 1) < 31$. The flattening results from spectral leakage contaminating phase estimates with out-of-band phase contributions which are constrained to the first branch of the inverse tangent function (10). Because the phase at high frequencies is generally random due to the absence of strong signal, phase unwrapping does not solve the problem. This bias can be demonstrated using synthetic data generated by Fourier interpolation with known subsample lags (Figure 4). We have found an acceptable trade-off between the advantages of the multitaper method (particularly quantitative error bars) and the need to reduce spectral leakage and resulting underestimation of the subsample lag, by using the average of the two lowest-order tapers to obtain the spectral values, while using the standard deviation of the six lowest-order tapers to estimate standard deviations on each phase point. We illustrate by showing the average cross-spectral phase in Figure 3d, calculated from the 0^{th} and 1^{st} order spectra represented by solid lines in Figure 3c. Error bars in Figure 3d were calculated using all six of the eigen-cross-spectra shown in 3c. We further downweight the most uncertain phase values by stretching the standard deviations of the multitaper estimates, σ_k, using the mapping

$$\sigma'_k = \tan(\sigma_k), \quad (-\pi/2 + \varepsilon < \sigma_k < \pi/2 - \varepsilon) \ , \tag{13}$$

where $\varepsilon = 0.01$, and removing phase points from the phase slope estimation if σ_k is outside of the range specified in equation (13). The phase slope and its standard deviation are estimated using the L_2 (least-squares) zero-intercept linear regression for the K usable data points with:

$$\frac{d\phi}{df} = \frac{\sum_{i=1}^{K} \phi_i/\sigma'_i}{\sum_{i=1}^{K} f_i/\sigma'_i} \tag{14}$$

$$\sigma_\beta(d\phi/df) = \frac{\left(\sum_{i=1}^{K} 1/\sigma'_i\right)^{1/2}}{\sum_{i=1}^{K} f_i/\sigma'_i}. \tag{15}$$

Because we employ a preliminary integer cross-correlation step that adjusts the traces to the nearest sample prior to invoking the cross-spectral phase slope method, we do not need to concern ourselves with phase unwrapping when estimating the slope of the cross-spectral phase. Event pairs which are sufficiently similar to be passed to the subsample lag estimator will have a maximum phase lag of $\pm\pi$.

The final suite of inter-event cross-correlation lags d is determined by summing the coarse, integer lag values l_β and the subsample lag estimates l_α. Total lag standard deviation estimates σ' are the quadrature sum of the coarse and fine lag standard deviations:

$$\sigma'_i = \sqrt{\sigma^2_{\alpha,i} + \sigma^2_{\beta,i}}. \tag{16}$$

Solving for Outlier-Resistant Pick Corrections

The desired N-vector of pick adjustments, $\vec{b}$, is the solution to

$$\mathbf{G}\vec{b} = \vec{d}, \tag{17}$$

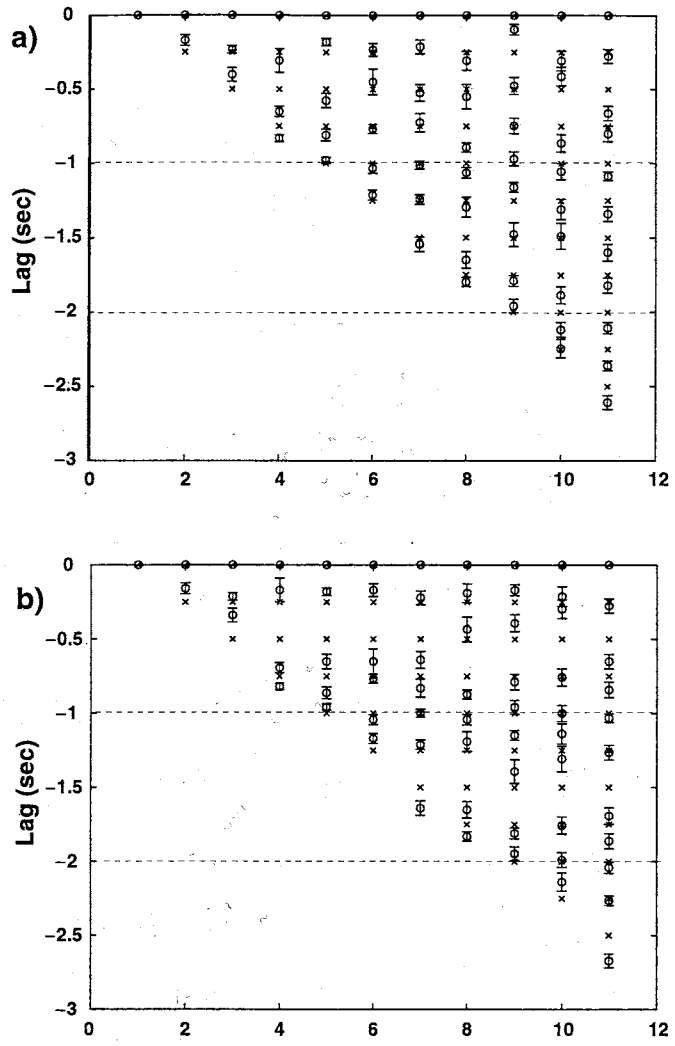


Figure 4. Example of phase slope flattening effect with higher order eigentapers. Using 32-sample synthetic seismograms whose true inter-event pick lags occur at 0.25 sample increments (indicated by crosses on both panels), we calculate subsample lags using the multiple eigentaper method. a) Automatically determined pick adjustments using only the two lowest-order eigentapers shows subsample lag estimates (circles) and associated 1σ errors (vertical error bars). Note that where the subsample lags are not properly estimated, there is a roughly random distribution of errors and these miscalculations fall largely within the estimated error bar for each lag. b) Subsample lag estimates calculated using eight lowest order eigentapers. Note that although the integer sample corrections have been properly applied, subsample shifts away from integer lag values are too small (circles tend to cluster about the integer lag values). From Aster and Rowe (2000).

where d is a M (up to $N(N\text{-}1)/2$-length) vector of weighted inter-event lags,

$$d_i = \frac{l_{\alpha,i} + l_{\beta,i}}{\sigma'_i} \quad (i \neq M). \tag{18}$$

the system matrix, **G**, is a weighted first-difference operator on $\vec{b}$ of the form:

$$\mathbf{G} = \begin{bmatrix} -1/\sigma'_{2,1} & 1/\sigma'_{2,1} & 0 & 0 & 0 & 0 & \dots \\ -1/\sigma'_{3,1} & 0 & 1/\sigma'_{3,1} & 0 & 0 & 0 & \dots \\ -1/\sigma'_{4,1} & 0 & 0 & 1/\sigma'_{4,1} & 0 & 0 & \dots \\ 0 & -1/\sigma'_{3,2} & 1/\sigma'_{3,2} & 0 & 0 & 0 & \dots \\ 0 & -1/\sigma'_{4,2} & 0 & 1/\sigma'_{4,2} & 0 & 0 & \dots \\ 0 & -1/\sigma'_{5,2} & 0 & 0 & 1/\sigma'_{5,2} & 0 & \dots \\ 0 & 0 & -1/\sigma'_{4,3} & 1/\sigma'_{4,3} & 0 & \dots & \dots \\ 0 & 0 & -1/\sigma'_{5,3} & 0 & 1/\sigma'_{5,3} & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \end{bmatrix}. \tag{19}$$

G is sparse; its $M \times N$ dimension results in $MN \propto N^3$ entries, of which only $2M \propto N^2$ entries are nonzero. This system sparseness can be exploited to reduce greatly computer storage and solution time. We parameterize **G** by two M-length index vectors, $\vec{A}^-$ and $\vec{A}^+$, which contain the row indices of the negative and positive entries for each constraint, and by an M-length vector, $\vec{\sigma'}$, containing the lag standard deviations. This storage scheme is capable of easily representing very large (tens of millions of elements) **G** systems within currently available workstation memory limits. Details of solving this system will be discussed below.

Clustering of similar waveform families

If many disparate earthquake families are included in the above system of equations, we may encounter the problem of matrix singularity or, at the very least, we will be making inappropriate comparisons among inconsistent waveforms. It then becomes necessary to decouple the system into sub-groups, each of which can successfully be solved for

consistency among related waveform correlation lags.

In a technique devised for obtaining relative hypocenter locations by applying cross-correlation-based traveltime differentials in the hypocenter solution problem, Waldhauser et al. (1999) addressed this problem by including a distance-weighting function in their singular value decomposition of the constraint system, so that inter-event waveform lags are weighted in the inversion based on proximity of preliminary hypocenter estimates. In cases where the distance from the seismic source to nearby stations is large compared to the geographic extent of the source region itself, consistent raypath parameters may be assumed, permitting the use of distance weighting to ensure that there is no inappropriate comparison of unrelated waveforms. A further advantage in applications such as those by Waldhauser et al. (1999) is that the network may contain many stations; preliminary misidentification of a few phases therefore will not seriously perturb the preliminary hypocenter locations on which the geographic proximity weighting is based.

Because we will be applying our technique not only to rather well-constrained and geographically extended systems of tectonic earthquakes, but also to the more problematic situations of limited and heterogeneous source regions such as geothermal fields and volcanic seismicity, we encounter additional difficulties (Rowe et al., 2000b,c). Seismic networks on volcanoes are often sparse, and the environment is invariably noisy. Additionally, signal quality can be poor due to high seismic attenuation of the volcanic pile. The confusing mix of tremor, wind noise, rockfall and (sometimes) glacier activity (e.g. Weaver and Malone, 1976, 1979) provide significant challenges to consistent phase identification for preliminary event locations. The resulting hypocenters can be significantly in error, particularly relative to the limited source dimensions of the volcanic seis-

micity. Although the recording environment in the case of a geothermal reservoir may be substantially better than the volcanic setting - reservoir sensors may be located in deep boreholes away from surface-related noise - the cost of borehole installations is high; therefore, such networks tend on the whole to be minimally configured for detection and location of microseismicity. In this setting a single mispick can have a large influence, and may result in appreciable hypocenter mislocation.

We further argue that in both volcanic and reservoir settings, events which do occur close to one another may have vastly different waveforms at any given station. In the case of a fractured reservoir, dimensions for intersecting joints and fractures may be quite small, and the failure induced on adjacent or intersecting features may exhibit dissimilar source functions (e.g., Phillips, 2000). Seismometers are generally located close enough to the source volume that minor differences in source mechanism or path can result in uncorrelatable waveforms among events located only a few meters apart. In a volcanic setting, brittle-failure volcano-tectonic (VT) events may occur on adjacent small fractures as the introduction of magmatic fluids perturbs the local pressure field and induces failure and hydrofracture; they may also occur in close proximity to dissimilar sources such as volcanic long-period (LP) and "hybrid" events (e.g., Chouet et al., 1994).

We therefore use waveform similarity-based clustering for preliminary classification and to ensure appropriate and stable pick lag solutions, with no constraints based on preliminary event location. Many candidate approaches for similarity clustering exist. Among those which have been employed successfully in other seismological applications are signal envelope cross-correlation (Carr et al., 1998), sonogram pattern recognition (Joswig, 1995) - sometimes referred to as "spectral fingerprinting" - and multi-station

median waveform cross-correlation (e.g. Aster and Scott, 1993). We have chosen waveform cross-correlation as our clustering criterion; the clustering may be selected based on correlation values for a single station, the median correlation value for a suite of stations, or another criterion best suited to the catalogue being evaluated. For the Redoubt dataset we cluster using waveform similarity for an extended envelope at a "master" station. Performing the clustering early in the analysis boosts the efficiency of our technique, as time and memory requirements for correlation and lag estimation on resulting sub-clusters decrease quadratically with the number of events.

An agglomerative, dendrogram-based hierarchical pair-group clustering algorithm has been chosen (e.g., Lance and Williams, 1967; Sneath and Sokal, 1973; Ludwig and Reynolds, 1988). Available clustering options include centroid mean, centroid median, single-link, complete-link, or flexible combinational weighting. Our technique was implemented following the method in the MATLAB-based seismic analysis package, MATSEIS (Harris et al. (1997). We outline the algorithm and its implementation for seismic waveform clustering below:

The $\vec{A}^{+}$, $\vec{A}^{-}$ event index vectors and the $\vec{\mathbf{k}}$ vector of inter-event cross-correlation maxima are used to build a dissimilarity matrix $\mathbf{K}$ upon which the clusterer may operate, where

$$\mathbf{K}_{i,j} = 1.001 - \mathbf{k}_m \quad . \tag{20}$$

The corresponding event indices are

$$i = \vec{A}^{+}_m \ , \ j = \vec{A}^{-}_m \tag{21}$$

The $\mathbf{K}_{i,j}$ may be viewed as a measure of inter-event distance in waveform similarity space for events i and j, where a value of ~0 equates with co-location and a value of ~1

represents infinite similarity distance. The algorithm addresses the entire suite of event pair distances, tracking the growing agglomerative hierarchy of family membership until all events are associated (Figure 5).

Table 1: Cluster combinational weighting parameters for different hierarchical clustering strategies. The number of entities in the *j*th and *k*th groups are represented by $t(j)$ and $t(k)$, respectively, and the number of entities in the combined (j,k) group is $t(j,k)$. (After Ludwig and Reynolds, 1988)

Strategy	α_1	α_2	β
Centroid (unweighted centroid)	$\frac{t(j)}{t(j,k)}$	$\frac{t(k)}{t(j,k)}$	$\frac{-t(j)t(k)}{t(j,k)}$
Centroid (weighted)/ median	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{4}$
Group mean/unweighted pair-grouping method	$\frac{t(j)}{t(j,k)}$	$\frac{t(k)}{t(j,k)}$	0
Flexible	0.625	0.625	-0.25

The algorithm begins by joining the two events *i* and *j* with the smallest inter-event dissimilarity value (equation (21)). Their inter-event dissimilarity distance is given by the $\mathbf{K}_{i,j}$ entry. The two rows and two columns, *i* and *j*, are replaced with a new vector $\mathbf{k}'$ whose entries are weighted by α_1, α_2 and β (Lance and Williams, 1967) (Table 1). For *n* events,

$$\text{for } m = 1{:}n \quad \mathbf{k}'_m = \alpha_1 \mathbf{K}_{i,m} + \alpha_2 \mathbf{K}_{j,m} + \beta \mathbf{K}_{i,j} \tag{22}$$

The matrix is again searched for minimum distance, and individual events may continue to be grouped in this manner until the shortest distance is found to belong either to a pre-existing cluster with an individual event, or two pre-existing clusters. The new combinational vector $\mathbf{k}''$ is derived using the linear combinational equation (Lance and

Williams, 1967) of the form:

$$\text{for } m = 1{:}n \quad \mathbf{k''}_m = \alpha_1 \mathbf{K}_{(i,h),m} + \alpha_2 \mathbf{K}_{(j,h),m} + \mathbf{K'}_{m,h}\,, \tag{23}$$

where the distance between the cluster **<i,j>** and the new entity **h** may be computed from the known distances $\mathbf{K}_{i,h}$ and $\mathbf{K}_{j,h}$, and the weighting parameters α_1, α_2 and β. These combinational steps are repeated, and the **K** matrix reduced, until all events have been associated into a single group, requiring a total of n-1 cycles for n events For the waveform similarity problem, we choose the "flexible" combinational weighting scheme. This weighting scheme has coefficients chosen so that the sum of the α_1, α_2 and β parameters equals one, which means that with successive joinings, the $\mathbf{K'}$ matrix values (distances between clusters) moves in a single direction, either continually contracting or continually expanding, with no reversals of direction that would cause problems for an automated system. Further, a choice of -0.25 for β is "non-distorting" of the similarity space, meaning that inter-cluster values tend not to exceed the extrema of the initial **K** matrix (e.g. Sneath and Sokal, 1973).

The final problem is determining at what point to terminate clustering in an automated manner. Harris et al. (1997) have obtained good seismicity clustering results using the technique of cophenetic correlation (Sneath and Sokal, 1973), wherein the distances between elements in the $\mathbf{K'}$ matrix are compared at each clustering step with the event distances in the original **K** matrix.

The cophenetic correlation for each pair-wise combination (of entities or multi-event groups) is calculated after each fusion step using the original **K** matrix and a cophenetic matrix $\mathbf{K^c}$. The cophenetic matrix begins as a duplicate of the original matrix, but with each successive grouping, entries in $\mathbf{K^c}$ associated with the clustering step are

replaced with the new $\mathbf{k}'$ values, determined through the equations (22) and (23). This reduces the similarity of $\mathbf{K^c}$ to the original $\mathbf{K}$ matrix. The ith value of the cophenetic correlation parameter (for the ith clustering step) is then:

$$C_i = \frac{\sum_{j=1}^{n}\sum_{k=1}^{n} \mathbf{K^c}_{j,k}\mathbf{K}_{j,k}}{\left[\sum_{j=1}^{n}\sum_{k=1}^{n} \mathbf{K}_{j,k}\sum_{j=1}^{n}\sum_{k=1}^{n} \mathbf{K^c}_{j,k}\mathbf{K^c}_{j,k}\right]^{\frac{1}{2}}} \quad , \qquad (24)$$

where j and k represent the entities being combined in the ith correlation step (either individuals or previously clustered groups). As larger groups are formed, the correlation between the $\mathbf{K^c}$ matrix and the original $\mathbf{K}$ matrix will continue to decrease, and so will overall values of C_i, although the function decrease is not necessarily monotonic. We target the largest drop in cophenetic correlation as the point immediately before which fusion should stop, as this represents the transition where the greatest leap in disparity between $\mathbf{K_c}$ and $\mathbf{K}$ occurs.

We illustrate the clustering technique with a 12-object example dendrogram and cophenetic correlation function in Figure 5. Each joining on the dendrogram represents the identification of "smallest distance" in similarity space for the entries in the reduced matrix at each fusion step. In this example objects 7 and 8 are most similar, so are joined as a pair by equation (21). The matrix is reduced through annihilation of rows and columns 7 and 8, with a replacement row and column added whose entries are the new relationship of cluster <7,8> to the remaining ten members, governed by equation (22). We next join objects 10 and 11, then 5 and 6, then 3 and 4. The fifth combinational step finds the smallest distance in the reduced matrix to lie between cluster <7,8> and object 9, so

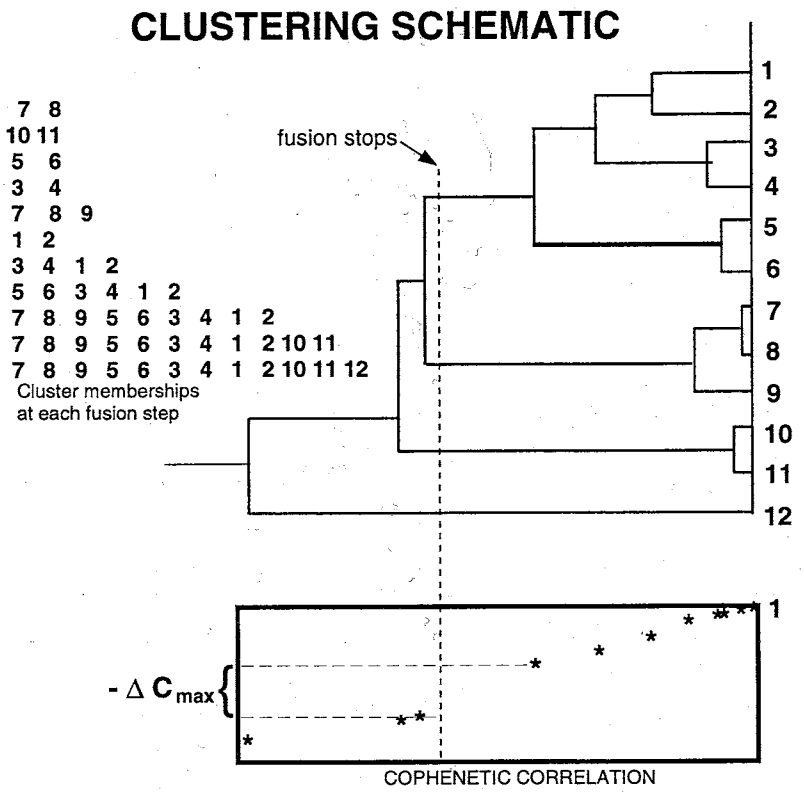


Figure 5: Schematic illustration of dendrogram-based, hierarchical pair-group clustering method. 5a) The dendrogram is built beginning by selection of the most similar event pair. Its combination yields a new single cluster entity, which is then compared against all other events. Subsequent joinings may be between two individual events, one event and a pre-existing cluster, or between two clusters, depending upon the values in the reduced similarity matrix. The cophenetic correlation parameter (plotted beneath the dendrogram) is calculated with each fusion step. Fusion continues until all events have been associated; the retroactive segregation cutoff is chosen as the step prior to the greatest drop in cophenetic correlation value.

these are also joined under equation (21). The process continues until step eleven when all entities are united. The point at which to stop fusion is made retroactively, using the maximum negative derivative of the cophenetic correlation function, *C*. At each step we have calculated a value for the cophenetic correlation, displayed beneath the dendrogram in Figure 5. The greatest drop in cophenetic correlation occurs at the fusion of cluster <5,6,3,4,1,2> with cluster <7,8,9>, which implies that in the hierarchy of this dendrogram, the greatest dissimilarity occurs at this fusion step. We therefore segregate the dataset into cluster memberships as defined immediately prior to this fusion step, leaving three clusters, <5,6,3,4,1,2>, <7,8,9>, <10,11>, and an orphan object <12>, which is not very similar to any of the others. Other automatic decision-making techniques exist, such as comparing the inter- and intra-cluster variances (e.g., Ludwig and Reynolds, 1988). We find, however, that the cophenetic correlation seems well suited to catalogues that divide robustly into unrelated groups of distinct waveforms, such as the discrimination of mine blasts (e.g. Carr et al., 1999a,b) from different locations, or teleseisms from different source regions.

Efforts to apply the cophenetic correlation method to catalogues of earthquakes exhibiting more continuous waveform variation, however, were less satisfactory (e.g., Rowe et al., 2000c). Under such circumstances the cophenetic correlation function becomes erratic, and a derivative-based termination of fusion on such a function is unreliable. We have therefore modified the algorithm so that we may instead select a similarity threshold, below which fusion stops. This approach yields a large number of small (doublet and multiplet) similarity groups and may be overly aggressive in cases where large general earthquake families are present. In our companion paper (Rowe et al., 2000a) we

use the cophenetic correlation approach, as it yields good success in terms of segregation of volcanotectonic and volcanic long-period waveforms among the Redoubt Volcano seismicity. The final cluster memberships under any segregation scheme depend strongly upon the length of correlation window and the degree of filtering; hence, some interactive testing on a random catalogue subset is advisable to determine such parameters.

Once a satisfactory division has been found, the catalogue is separated into corresponding clusters. Individual phases are cross-correlated among member earthquakes of each cluster. Within each cluster, optimal pick adjustment may be found to be critically sensitive to the correlation window length; hence, the algorithm applies cross-correlation for each phase using a suite of correlation windows. Criteria for automatically selecting the most desirable correlation for each phase are the mean cross-correlation value for all constraints and the mean (or median) of the estimated lag standard deviations (equation (14)) among the different trials. After the optimum lags have been chosen, we solve for consistent pick adjustments for all member events within a cluster.

Solving the systems of constraints for consistent pick lags

Straightforward linear approaches to solving equation (17) for a least-squares residual (L_2) solution include Cholesky factorization or other techniques of solving the normal equations, or through singular-value decomposition (e.g., Press et al., 1996). Such methods, however, require calculation of the (non-sparse) $\mathbf{G}^T\mathbf{G}$ (which contains $M^2 \propto N^4$ entries) or other large intermediary objects, which eliminate computational and storage advantages associated with our ($\vec{A}^+$, $\vec{A}^-$) storage of the sparse $\mathbf{G}$. Additionally, the L_2 solution has the undesirable property of being strongly perturbed by outliers (e.g. Parker and McNutt, 1980; Shearer, 1998).

We instead solve equation (17) by implementing an iterative Polak-Ribiere conjugate gradient minimization (Polak, 1971; Press et al., 1996) formulated to operate efficiently with the $(\vec{A}^{+}, \vec{A}^{-})$ sparse storage scheme. This is also implementable for the more robust minimum 1-norm residual (L_1) solution. The functional to be minimized is

$$f = \mu^{(1)} = \sum_{i=1}^{M} \frac{|d_i - d_{i,\,\mathrm{pred}}|}{\sigma_i}; \tag{25}$$

and the gradient of f at a solution space point, $\vec{x}$, needed for the conjugate gradient technique is

$$\vec{\nabla} f = \frac{\mathrm{sgn}(\mathbf{G} \cdot \vec{x} - \vec{d})}{\sigma}, \tag{26}$$

where the sgn function returns 1 if the argument is positive, -1 if the argument is negative, and 0 if the argument is zero. Although it has superior resistance to outliers, implementation of the L_1 residual minimization becomes problematic when any of the residuals becomes too small, as the derivative function becomes discontinuous. We have successfully addressed this difficulty by modifying the misfit function for values of f which lie within the region $-\varepsilon < 0 < \varepsilon$ for small ε:

$$\text{if } |f_i| > \varepsilon, \quad f_i = \frac{|d_i - d_{i,\,\mathrm{pred}}|}{\sigma_i}, \tag{27}$$

and $\vec{\nabla} f$ is as described in equation (26).

$$\text{if } |f_i| \le \varepsilon, \quad \frac{\langle d_i - d_{i,\,\mathrm{pred}} \rangle^2}{2\sigma\varepsilon} + \frac{\varepsilon}{2}, \tag{28}$$

$$\text{and} \quad df_i = \frac{\mathrm{sgn}(d_i - d_{i,\,pred})}{\sigma\varepsilon} \tag{29}$$

This modification has a theoretical drawback insofar as the smallest misfit we may obtain is $\varepsilon/2$, as opposed to zero; however, this poses no practical difficulty. We are currently obtaining satisfactory results using a value of $\varepsilon = 0.1$. Calculation of the solution probability (outlined below) may be done by re-computing f with the exact L_1 formula-

tion, although this will not be the true minimum because of our approximation for small f_i.

From a probabilistic viewpoint, the L_1 solution is the maximum likelihood under the assumption of exponentially distributed data errors, described by

$$P^{(1)}(x) = \frac{1}{\sigma} e^{-|x-m|/\sigma} , \tag{30}$$

rather than the shorter-tailed Gaussian (L_2) data errors, described by

$$P^{(2)}(x) = \frac{1}{(\pi\sigma)^{1/2}} e^{-(x-m)^2/\sigma^2} . \tag{31}$$

Parker and McNutt (1980) describe the statistics of $\mu^{(1)}$ under an assumption of Gaussian data errors, which we invoke as a useful quality-of-fit measure to assess whether the relative lags estimated by our L_1 solutions form a consistent set of first-difference constraints on $\vec{b}$ (equation (17)). The L_1 analogue to the L_2 (χ^2) q statistic for M = K-N degrees of freedom is approximated by a third-moment expression for the probability that a greater value of $\mu^{(1)}(M)$ than the observed one (equation (25)) could have occurred:

$$q(f, M) = P(\mu^{(1)}(M) > x(f, M)) = P(x) - \frac{\gamma}{6} Z^{(2)}(x) . \tag{32}$$

The first term on the right-hand side of Eq. 22 is the cumulative probability integral

$$P(x) = \frac{1}{\sigma_1 (2\pi)^{1/2}} \int_{-\infty}^{x} e^{-t^2/\sigma_1^2} dt \tag{33}$$

for a zero-mean Gaussian distribution with the variance of $\sigma_1 = \sigma^2(\mu^{(1)})$, where

$$\sigma_1^2 = (\, 1 - 2/\pi \,) M . \tag{34}$$

The second term on the right-hand side of Eq. 22 is proportional both to the skewness, γ, of $\mu^{(1)}$:

and to

$$\gamma = \frac{2 - \pi/2}{(\, \pi/2 - 1 \,)^{3/2} M^{1/2}} , \tag{35}$$

where

$$Z^{(2)}(x) = \frac{1}{(2\pi)^{1/2}} (x^2 - 1) e^{-x^2/2} \tag{36}$$

and $$x = \frac{f - \bar{\mu}}{\sigma_1}. \tag{37}$$

The mean value of $\mu^{(1)}$ is $$\bar{\mu} = (2/\pi)^{1/2} M . \tag{38}$$

Although the L_1 solution is tolerant of outliers, we have found that we can nevertheless improve the L_1 solution by conservatively rejecting correlation outliers. The first step is to discard lag constraints whose cross-correlation maxima are sufficiently poor that there is little likelihood of constructive contribution to the solution. We have adopted an *a priori* threshold of 0.8 in our example application; any constraints which fail to meet this minimum are rejected prior to the first attempt at solving the system. This tolerance will vary depending upon the quality of the catalogue being addressed and the desired similarity threshold in the application. Preliminary clustering helps to ensure that disparate families of earthquakes are not being compared, but outliers and poorly correlating events resulting from correlation cycle skips, excessive noise or grossly inaccurate initial picks may still remain.

To eliminate the influence of these outliers, we first calculate the L_1 solution to equation (17) and its misfit measure, f (equation (25)), using the full constraint and data set for the cluster (minus the *a priori* rejections). If there are data outliers or the system is otherwise highly inconsistent, a large value of y will produce a highly unlikely value of $q(f,M)$ (close to zero). If q is very small, we successively cull constraints and corresponding data from the system, using a binary search mechanism. The data misfit vector is sorted and the constraints corresponding to the worst half of the misfit estimate are discarded. Discarding, rather than downweighting, the highest misfit constraints assures correct probability calculations for subsequent solutions by appropriately adjusting the degrees of freedom of

the system. The reduced system is solved, and the value of $q(f^i,M^i)$ is re-calculated for the ith bisection step under the new degrees of freedom. If this value is too good ($q(f^i,M^i) >$ 0.02), we assume that too many constraints have been discarded and we restore a portion of them. We re-compute $q(f^i,M^i)$ and restore or discard constraints again, as appropriate. This process generally converges to a satisfactory value of $q(f^i,M^i)$ within ten steps, but depends upon pre-determined thresholds for convergence and bisection step size parameters. After convergence has been achieved, we obtain a final pick adjustment solution for the reduced system of M' constraints, and calculate 1-sigma error bars for each element of the final solution via Monte Carlo propagation of Gaussian data errors. A *post priori* zero mean constraint is applied to the final set of pick adjustments.

Figure 6 shows a synthetic illustration of the solution process for an m = 6-event synthetic cluster, initially constrained by a full set of $M = (6)(5)/2 = 15$ inter-event lag estimates and 1 zero-mean constraint. The true pattern of pick adjustments was chosen arbitrarily to be a zero-mean half-period sine function with an amplitude of 1 time unit. The 15 first-difference data points were randomized by adding a Gaussian error term with a standard deviation of 0.2 time units. Outliers were introduced to the system by adding large random terms to data points 3 and 7.

Figures 6a and 6b show the L_1 solution and data fit for the entire data and constraint set, where the recovery of true pick adjustments has been skewed by the data outliers, and the probability of a worse misfit is $q \approx 0$ to single-precision. After automatically rejecting the two system constraints with the largest residual contributions, as described above, and re-solving the problem, we obtain a revised solution (Figure 6c) with an acceptable data misfit (Figure 6d) of $q(9.22, 8) \approx 0.06$ and an L_1 misfit improvement

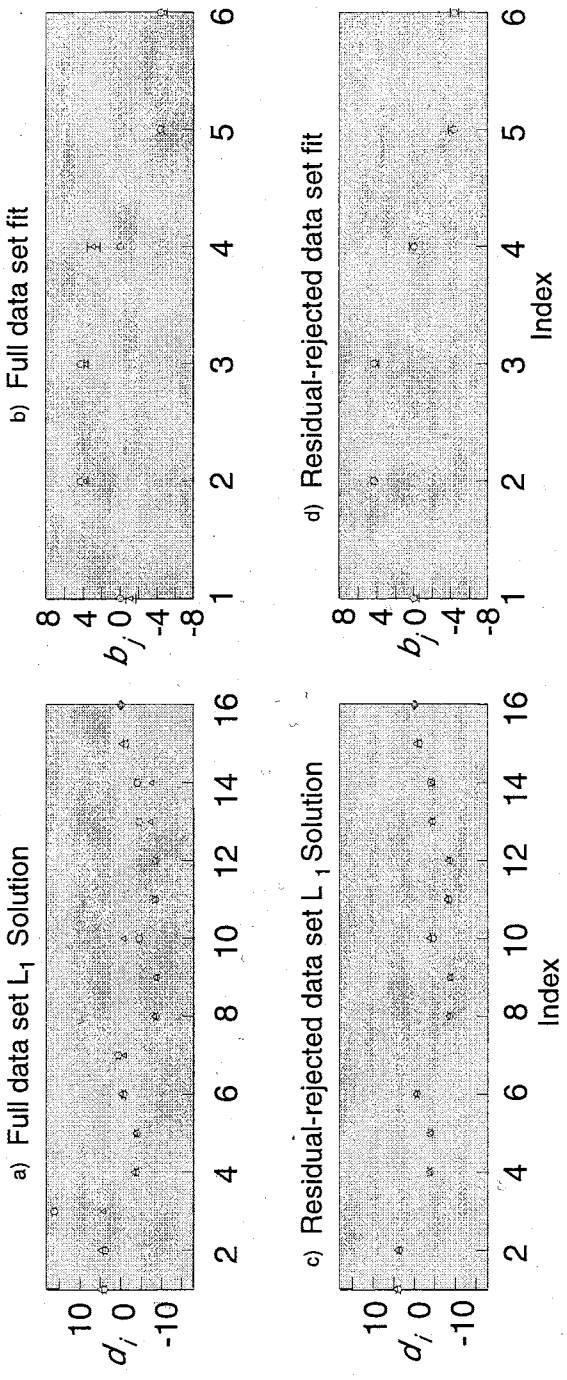


Figure 6. Improved solution robustness using the L_1 residual minimization and iterative residual-based system constraint rejection. *a)* Full data set solution (triangles) with Monte Carlo estimated standard deviations, compared to the true model (circles). b) Predicted data are shown as triangles and the actual data is shown as circles accompanied by standard errors. c) Reduced dataset with outliers removed by residual-based rejection. d) Refined solution following removal of outliers.

between solution and true model,

$$f' = \sum_{i=1}^{M'} \frac{|x_i - x_{i,\text{true}}|}{\sigma_1}, \tag{39}$$

of 59% with generally tighter 1-sigma error bars.

Absolute versus relative locations

Introduction of the new picks for each cluster provides precise relative event relocations within clusters, but the question of improved inter-cluster locations is not addressed in this fashion. Within any given cluster it is commonly observed that analyst picks do not always scatter about a zero mean; they are often systematically late in instances of low signal/noise. The resulting adjusted picks for a particular phase may therefore exhibit a significant bias, and an assumption of mean cluster centroid is no longer appropriate.

Our approach to addressing inter-cluster pick adjustments avoids any dependence upon preliminary location parameters. Relative pick lags within clusters are estimated as outlined above, adjusting the phase picks accordingly. Waveforms for each phase within each cluster are then aligned on their adjusted picks and stacked (Figure 7). Each stack is then treated as a composite earthquake trace for the cluster. Ensembles of stacked seismograms are then cross-correlated and relative pick lags determined between the composite earthquakes using the L1-norm conjugate gradient solver as before. The resulting inter-cluster lags are used to adjust mean picks within each cluster. In this way the very tightly constrained relative adjustments for intra-cluster associations are preserved, with no risk of mistakenly including uncorrelated events through preliminary mislocation.

In Figures 7a, 7b and 7c we show preliminary and repicked waveform alignments

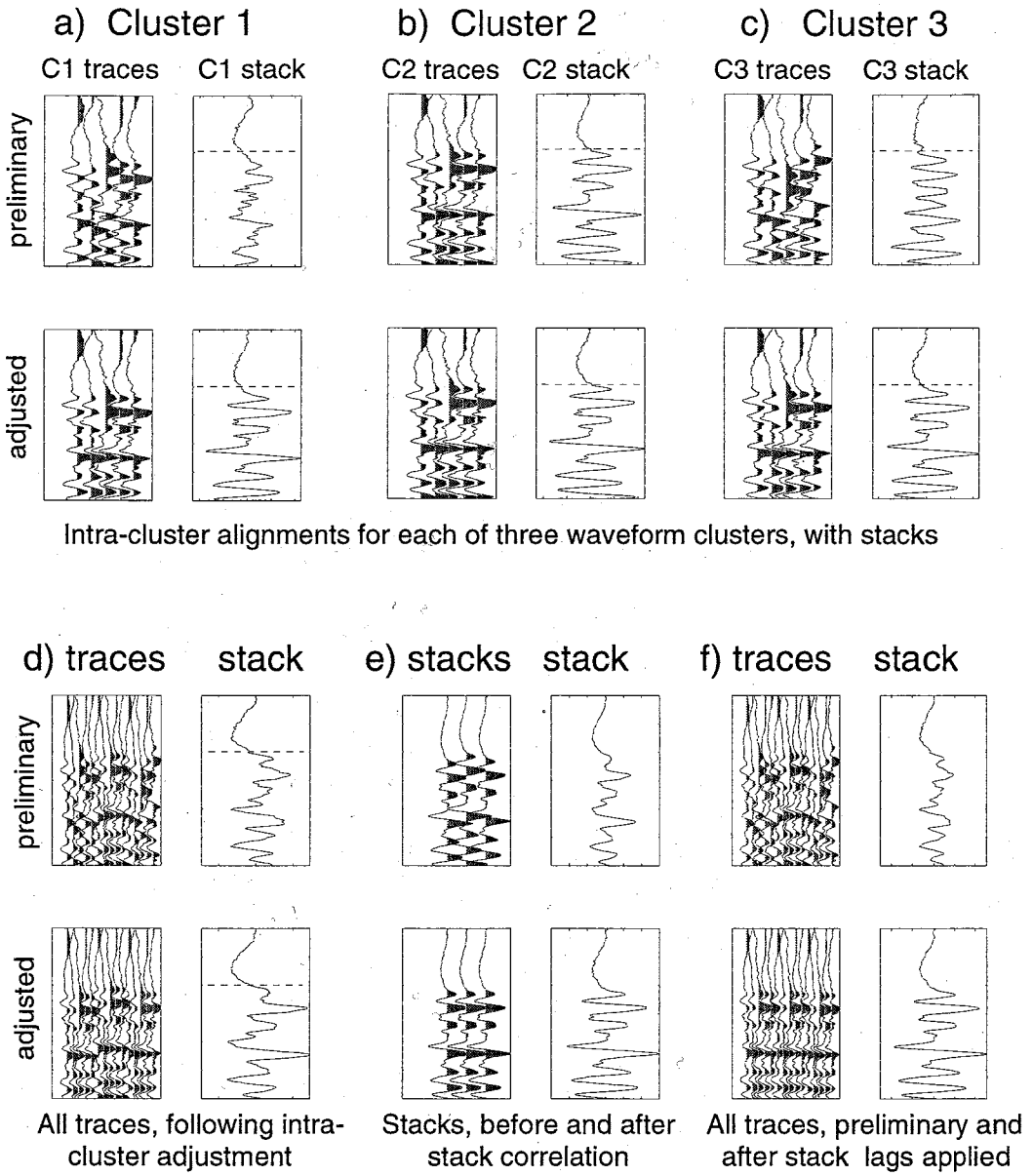


Figure 7: Hierarchical clustering, lag adjustment and stacking method. a)-c) illustrate three hypothetical clusters of five events each. Upper panels show traces aligned on preliminary picks and associated waveform stack; lower panels show traces aligned on adjusted picks, with associated stack. Horizontal dashed lines indicate pick times on the stacked trace in each. d) Clusters from a), b) and c) combined to show initial scatter (upper panel) and relationships among intra-cluster adjustments for the three clusters. e) Stacks from clusters of a) b) and c) showing preliminary alignment (upper panel) and appropriately shifted stacks, following hierarchical stacking and cross-correlation (lower panel). f) Three clusters showing initial misalignments (upper panel) and final, corrected alignments (lower panel) after both intra-cluster and inter-cluster lags have been applied.

for events from three synthetic clusters, recorded at the same station. In this example we have used the same waveforms for each example cluster, with different initial pick errors. Although within each of the aligned clusters the new picks are consistent, note that the mean adjusted pick occurs at somewhat different times on the resulting alignment stack. This is an artifact of the differing preliminary pick distributions among each of the clusters. Such biases will result in the mislocation of inter-cluster mean centroids. Figure 7d illustrates the problem which results if we assume that preliminary cluster centroids are accurate: in the upper panels we show all member events of the three clusters, aligned on their preliminary picks and then stacked, and in the lower panels we show the traces aligned on their revised intra-cluster picks, and the resulting stack. Although the waveforms have clearly been well-aligned for intra-cluster consistency, a serious misalignment is exhibited among the three clusters.

If we subsequently cross-correlate the stacked waveforms and determine relative lags among the stacks (Figure 7e), we can then apply the additional pick correction to each of the member traces and adjust all events by the inter-cluster centroid shifts (Figure 7f). This hierarchical, bi-level clustering and stacking approach still does not address the question of overall analyst bias for an entire catalogue, but any remaining artifact could be handled through individual station corrections determined through JHD or other joint location methods. One drawback to such an approach is that the picks themselves are not corrected for the final mean catalogue bias, and the adjustment dies not separate picking bias from actual traveltime residuals at the receivers. Another approach to addressing any remaining pick bias may be to apply a computer auto-picking algorithm to the final stacked waveform (for example, the second-order adjusted stacked trace in the lower right-hand panel

of Figure 7e), or to manually adjust the picks on the second-order stacks.

We present our autorepicking and subsequent relocation for the seismicity associated with the 1989-1990 eruption of Redoubt Volcano, Cook Inlet, Alaska, in a companion paper (Rowe et al., 2000a).

SUMMARY

We have developed an automatic, adaptive algorithm for adjusting phase picks for consistency among similar events within large digital seismic waveform catalogues. Innovations include automatic, adaptive cross-coherency and polarization filtering, eigenspectral methods for estimating subsample phase lags and appropriate, dimensionally consistent lag standard deviations. After initial cross-correlation, the catalogue is clustered using a hierarchical, dendrogram-based pair-group classification scheme with segregation based upon the cophenetic correlation function. Resulting clusters are solved for consistent intra-cluster pick lags using an L1-norm minimizing, outlier-resistant, iterative, conjugate gradient method formulated to minimize memory and computation requirements. Intra-cluster seismograms are aligned on the zero-mean adjusted repicks, then stacked to provide a composite waveform. Ensembles of stacked seismograms are then cross-correlated among clusters to determine inter-cluster pick lag adjustments; this corrects for possible cluster centroid biases and provides for consistent inter-cluster pick (and location) relationships within the catalogue, with no dependence upon preliminary hypocenter and traveltime parameters.

In a companion paper (Rowe et al., 2000a) we apply our method to seismicity associated with the 1989-1990 eruption sequence of Redoubt Volcano, Alaska, to demonstrate earthquake location improvements which may be possible with heterogeneous

datasets recorded under problematic circumstances. The method is successful in segregating and properly associating volcano-tectonic and volcanic long-period earthquakes, and can provide meaningful fine adjustments to spatial and temporal seismicity trends, which might otherwise be misinterpreted due to preliminary location biases.

ACKNOWLEDGEMENTS

This work was supported under a grant from Sandia National Laboratories, Albuquerque, New Mexico, by Niitsuma Laboratories, Tohoku University, Sendai, Japan, and by National Science Foundation Office of Polar Programs grant #9419267.

REFERENCES

Aster, R.C. and J. Scott, Comprehensive characterization of waveform similarity in microearthquake data sets, *Bull. Seismol. Soc. Am. 83,* 1307-1314, 1993.

Aster, R.C. and C.A. Rowe, Automatic phase pick refinement and similar event association in large seismic data sets, in *Advances in Seismic Event Location*, edited by C. Thurber and N. Rabinowitz, pp. 231-263, Kluwer, Amsterdam, 2000.

Carr, D., Young, C., Harris, J., Aster, R., Zhang, X., Cluster Analysis for CTBT Seismic Event monitoring (abstract), *Seismol Res. Lett. 70*, 227-228, 1999.

Deichmann, N. and M Garcia-Fernandez, Rupture geometry from high-precision relative hypocentre locations of microearthquake clusters, *Geophys. J. Int. 110*, 501-517, 1992.

Digby, P.G.N. and R.A. Kempton, *Multivariate Analysis of Ecological Communities*, Chapman & Hall, London, UK, 1987.

Dodge, D.A., G.C. Beroza and W.L. Ellsworth, Foreshock sequence of the 1992 Landers, California earthquake and its implications for earthquake nucleation, *J. Geophys.*

Res. 100, 9865-9880, 1995.

Fehler, M.C., L.S. House and H. Kaieda, Determining planes along which earthquakes occur: Method and application to earthquakes accompanying hydraulic fracturing, *J. Geophys. Res.*, 92: 9407-9414, 1987.

Fehler, M., W.S. Phillips, R. Jones, L. House, R. Aster and C. Rowe, A method for improving relative earthquake locations, *Bull. Seismol. Soc. Amer. 90*, 775-780, 2000.

Fremont, M.-J. and S. D. Malone, High precision relative locations of earthquakes at Mount St. Helens, Washington, *J. Geophys. Res.* 92: 10,223-10,236, 1987.

Gillard, D., A.M. Rubin and P. Okubo, Highly concentrated seismicity caused by deformation of Kilauea's deep magma system, *Nature 384*, 343-346, 1996.

Goldstein, P., D. Dodge, M. Firpo, and S. Ruppert, What's new in SAC2000? Enhanced processing and database access, *Seismol. Res. Lett. 69*, 202-205, 1998.

Got, J.-L., J. Frechet and F.W. Klein, Deep fault plane geometry inferred from multiplet relative relocation beneath the south flank of Kilauea, *J. Geophys. Res. 99*, 15,375-15,386, 1994.

Harris, M. and C. Young, MatSeis: a seismic GUI and tool-box for MATLAB, *Seismol. Res. Lett. 68*, 267-269, 1997.

Ito, A., High resolution relative hypocenters of similar earthquakes by cross-spectral analysis method, *J. Phys. Earth 33,* 279-294, 1985.

Jones, R.H. and R.C. Stewart, A method for determining significant structures in a cloud of earthquakes, *J. Geophys. Res. 102*, 8245-8254, 1997.

Joswig, M., Automated classification of local earthquake data in the BUG small array,

Geophys. J. Int. 160, 262-285, 1995.

Kissling, E., Geotomography with local earthquake data, *Rev. Geophys. 26*, 659-698, 1988.

Lance, G.N. and W.T. Williams, A general theory for classificatory sorting strategies 1. Hierarchical systems., *Comp. J. 10*, 271-276, 1967.

Lee, W.H.K., The Xdetect program, a course on P-based seismic networks in *U.S. Geol. Surv. Open File Rep. 92-441,* edited by W.H.K. Lee and D.A. Dodge, pp. 138-151, 1992.

Ludwig, J.A. and J.F. Reynolds, *Statistical Ecology, a Primer on Methods and Computing,* John Wiley & Sons, New York, NY., 1988.

Park, J., C.R. Lindberg and F.L. Vernon III, Multitaper spectral analysis of high-frequency seismograms, *J. Geophys. Res. 92*, 12,675-12,684, 1987

Parker, R. and M. McNutt, Statistics for the one-norm misfit measure, *J. Geophys. Res. 85*, 4429-4430, 1980.

Phillips, W.S., L.S. House and M.C. Fehler, Detailed joint structure in a geothermal reservoir from studies of induced microearthquake clusters, *J. Geophys. Res. 102*, 11,745-11,763, 1997.

Phillips, W.S., Precise microearthquake locations and fluid flow in the geothermal reservoir at Soultz-sous-Forets, France, *Bull. Seismol. Soc. Am.* 90, 212-228, 2000.

Poupinet, G., W.L. Ellsworth and J. Frechet, Monitoring velocity variations in the crust using earthquake doublets: An application to the Calaveras Fault, California, *J. Geophys. Res. 89*, 5719-5731, 1984.

Press, W.H., B.P. Flannery, S.A. Teukolsky and W.T. Vetterling, *Numerical Recipes in C*,

Cambridge University Press, New York, 1989.

Pujol, J., Joint hypocentral location in media with lateral velocity variations and interpretation of the station corrections, *Phys. Earth Planet. Inter. 75*, 7-24, 1992.

Rowe, C.A., R.C. Aster, M.C. Fehler, W.S. Phillips and P.R. Kyle, Hypocenter Relocation in Tectonic and Volcanic Seismic Data Sets Using Automatic phase Correlation: Application of Cross-correlation, Signal Envelope and Principal Components Methods to Improve Pick Consistency, *EOS, Trans., Amer. Geophys. U. 79,* F588, 1998.

Rowe, C.A. and R.C. Aster, Application of automatic, adaptive filtering and eigenspectral techniques to large digital waveform catalogs for improved phase pick consistency and uncertainty estimates, *EOS, Trans., Amer. Geophys. U. 80*, F660, 1999.

Rowe, C.A., R.C. Aster, B. Borchers and C.J. Young, Relocation of seismicity associated with the 1989-1990 eruption of Redoubt Volcano, Alaska. Part I: An automatic, adaptive algorithm for refining phase picks in large seismic data sets, *J. Geophys. Res.*, in prep, 2000a.

Rowe, C.A., R.C. Aster, B. Borchers and C.J. Young, Relocation of seismicity associated with the 19889-1990 eruption of Redoubt Volcano, Alaska, Part II: Application to Redoubt eruption seismicity, *J. Geophys. Res.*, in prep, 2000b.

Rowe, C.A., R.C. Aster, W.S. Phillips, R.H. Jones, B. Borchers and M.C. Fehler, Relocation of induced microseismicity at the Soultz geothermal reservoir using automated, high-precision repicking. *Pure Appl. Geophys.*, submitted, 2000c.

Rubin, A. M., D. Gillard and J.-L. Got, A reinterpretation of seismicity associated with the January 1983 dike intrusion at Kilauea Volcano, Hawaii, *J. Geophys. Res. 103*, 10,003-10,015, 1998.

Rubin, A.M., D. Gillard and J.-L. Got, Streaks of microearthquakes along creeping faults, *Nature 400*, 635-641, 1999.

Shearer, P.M., Improving local earthquake locations using the L1 norm and waveform cross correlation: Application to the Whittier Narrows, California, aftershock sequence, *J. Geophys, Res. 102*, 8269-8283, 1997.

Shearer, P.M., Evidence from a cluster of small earthquakes for a fault at 18 km depth beneath Oak Ridge, Southern California, *Bull. Seismol. Soc. Am. 88*, 1327-1336, 1998.

Slunga, R., S.T. Rognvaldsson and R. Bodvarsson, Absolute and relative locations of similar events with application to microearthquakes in southern Iceland, *Geophys. J. Int. 123*, 409-419, 1995.

Sneath, P.H.A. and R.R. Sokal, *Numerical Taxonomy*, Freeman, San Francisco, CA, 1973.

Thomson, J.D., Spectrum estimation and harmonic analysis, *Proc. IEEE 70*, 1055-1096, 1982.

Waldhauser, F. W.L. Ellsworth and A. Cole, Slip-parallel seismic lineations on the northern Hayward Fault, California, *Geophys. Res. Lett. 26*. 3525-3528, 1999.

Weaver, C. S. and S. D. Malone, Mt. St. Helens seismic events: Volcanic earthquakes or glacial noise, *Geophys. Res. Lett. 3*, 197-200, 1976.

Weaver, C. S. and S. D. Malone, 1979, Seismic evidence for discrete glacier motion at the rock-ice interface, J. *Glaciol. 23*, 171-183.

Wiemer, S. and M. Wyss, Minimum magnitude of completeness in earthquake catalogs: Examples from Alaska, the western United States and Japan, *Bull. Seismol. Soc. Am. 90*, 859-869, 2000.

Appendix A1: CCHAR: Clustering, Correlation-based Hierarchical Auto-adaptive Re-picker

A1.1. Overview:

CCHAR and ancillary tools are intended to be available and useful to anyone in the seismological community who is interested. This is a work in progress; updates and improvements will be posted on the CCHAR websites whenever possible (**http://www.ees.nmt.edu/Geop/cchar.html** and **http://www.geology.wisc.edu/~char/**). Following is a set of user guidelines which may help to acquaint users with the package. In addition to this user guide, a test dataset has been put onto the CD-ROM at the back of this thesis, which will permit users to verify that **CCHAR** is running correctly for them.

There are several steps required to run **CCHAR** and get meaningful results; these can be roughly divided into the following stages:

Grooming the catalogue

Preliminary cross-correlation

Determining the clustering/segregation choices based on preliminary correlation

Invoking the clusterer, dividing the catalogue

Re-correlating with optimal intra-cluster parameters for each station in a cluster

Solving for consistent intra-cluster pick lags

Stacking intra-cluster traces, cross-correlating the stacks for inter-cluster corrections

Correcting the picks, building pickfiles or other needed results.

CCHAR is a combination of c-shell scripts, C-programs, licensed (and unlicensed) proprietary and non-proprietary software/shareware. There are also some useful associated Matlab scripts and functions for graphical display of results. Any of the pieces may be used for various purposes independently; the scripts are the glue that bind the sequence together and these will necessarily be changed and adapted according to user needs.

The C code was developed through interactive and incremental coding and testing within Matlab. Although an effort was made to write the code in ANSI-standard C, complete platform independence has not yet been achieved. This code, as it stands, runs on Sun Solaris 2.X systems. Dependence on SUN routines has been kept to a minimum; however, we employ the SUN random number generator, **random** and we currently invoke the SUNmath library for the **log2** function, which is used in allocating memory for some of the numerical sorting requirements. Additionally, inclusion of the standard math library may provide different functions which are implemented differently on other platforms. Modifying for general portability to LINUX and HP-UNIX as well as other platforms is planned.

The C codes included in the package currently rely on the presence of certain software libraries:

SAC: Because **CCHAR** currently operate on **SAC** format traces, the **sac.h** include file is needed. This file, which dictates the header structure for **SAC** traces, is included in the source code directory for the correlator. **SAC** is a standard seismological digital waveform format and is universally used among earthquake seismologists. Robust conversion rou-

tines such as **ah2sac**, **suds2sac**, **seed2sac** and **segy2sac** may be obtained from IRIS for converting **AH**, **SUDS**, **SEED** and **SEG-Y** (and so on) traces into **SAC** format; visit **http://www.iris.washington.edu** and search the **Seismic Software Library** for **fcu.tar** under the **Conversion Programs** heading. Please note that some data formats may lack header values for such parametric data as picks and event origins, so these may not be communicated into the **SAC** headers upon conversion. Some formats, such as **SEG-Y**, may have such header information as shot times, which do not have corresponding header locations in **SAC**, although they may be stuffed into another slot such as **user5**. To obtain **SAC**, visit the **SAC** homepage at **http://www.ep-es.llnl.gov/sac**

Numerical Recipes: **CCHAR** relies upon Numerical Recipes (**NR**) C routines including **jacobi.c**, **sort.c**, **sort2.c** and **indexx.c**. These in turn depend upon the **NR** include file, **nrutil.h**. Distribution of these subroutines is restricted under the licensing agreement for **NR** codes. Users whose systems are licensed for **NR** C codes may run **CCHAR** without modifying it to accept substitute routines. **NR** may be easily obtained through almost any bookseller (i.e. amazon.com) and is great to have around for a variety of computational needs, rather than re-inventing the wheel. The book itself provides very helpful explanations for many aspects of mathematical programming, including theoretical foundations. **NR** is also available in FORTRAN. A word of warning regarding **NR** C codes: Because they were originally developed in FORTRAN and subsequently modified to C, they contain some FORTRAN-isms, the most problematic being an indexing standard for *n*-length vectors which begins with 1 and goes to *n*, as opposed to the "**i=0;i<n;i++**" style common in C programming. **CCHAR** routines have been adapted to accept this convention where

necessary, so if non-proprietary substitutions are made for **CCHAR NR** calls, care should be taken regarding indexing. This also has ramifications for dynamic memory allocations for some of the arrays found in the driver programs **corr_sac, run_cgsolve** and **run_cluster**. As a rule, an extra element should always be allocated if these arrays are altered.

FFTW: Because of the requirement to calculate thousands to tens of millions of Fourier transforms in a single run of **corr_sac**, a new FFT routine was incorporated into **CCHAR**, which reduces the computation time significantly. This package, called **FFTW** ("Fastest Fourier Transform in the West") was developed at the Massachusetts Institute of Technology and is available free of charge at **http://www.fftw.org. FFTW** performs real, complex, multidimensional and parallel transforms with optimal speed on series whose lengths may be any combination of small prime numbers. The software also devises transforms on the fly to make optimal use of system architecture and resources, further enhancing performance. If your system does not have **FFTW** installed, you will need to modify all **fftw** and **rfftw** calls, remove the **create_plan** and **destroy_plan** calls and the structure declarations **fftw_real** and **fftw_complex**, as well as modify the **makefile**, to incorporate FFT routines already in residence on your machine. I strongly recommend obtaining and installing the **FFTW** package, as it will also enhance the performance and flexibility of other computational tasks you want to do that require Fourier transforms.

Multitaper routines: The multitaper calculation routines incorporated into **CCHAR** were obtained from Bell, B., Percival, D.B. and Walden, A.T. "Calculating Thomson's Spectral

Multitapers by Inverse Iteration", J. Comput. and Graph. Stat., 1993. These have been folded into the code so that tapers of the proper length are automatically pre-computed for any chosen correlation window. No special libraries are needed for the user, but as "imported" software they are hereby given due credit.

A1.2: Grooming the catalogue

CCHAR obtains the needed parametric information from the trace headers. **SAC** header format and storage locations can be found in the **SAC** manual. It can also be found interactively within **SAC** in the online help via "**help user_manual_part6**," as well as by visiting the **SAC** website, **http://www-ep.es.llnl.gov/sac**.

A1.2.1 SAC Header Variables

The following parameters will be expected by **CCHAR**, and should be stored in the indicated header positions:

Value	Header Variable
sample rate	**delta** (i.e. 100 Hz=.01)
trace beginning	**b** (almost always zero)
initial P pick	**a** (generally in seconds after b)
initial S pick	**t0** (also in seconds after b)
year trace begins	**nzyear** (4 digits in sac2000)
Julian day trace begins	**nzjday**
hour trace begins	**nzhour**
minute that trace begins	**nzmin**
second that trace begins	**nzsec**
millisecond trace begins	**nzmsec**
hypocenter latitude	**evla** (usually decimal degrees)

Value	Header Variable
hypocenter longitude	**evlo** (" " ")
hypocenter depth	**evdp** (usually km)

Other things may be in the headers, such as station information, event origin and so forth, but the program doesn't specifically need them in its present incarnation. Particular header variables should not be assigned by the user, as they will be pre-empted by **CCHAR**:

Value	Header Variable
Intra-cluster adjusted P pick	**t1**
Intra-cluster P pick standard deviation	**user1**
Intra-cluster adjusted S pick	**t2**
Intra-cluster S pick standard deviation	**user2**
Inter-cluster (stacked) final adjusted P	**t3**
Inter-cluster (stacked) P pick st.dev.	**user3**
Inter-cluster (stacked) final adjusted S	**t4**
Inter-cluster (stacked) S pick st.dev.	**user4**

A1.2.2 Data Organization

Picks are corrected using station-common gathers, so one station is addressed per run. **CCHAR** expects traces to be named identically for single-component data, and consistently for multicomponent data. Files are organized so that each event has a directory, which possesses all traces for that event. **CCHAR** operates on the files by searching through a **control file**, each line of which points to one trace. For example, in each of several event directories we have a common trace named **BGEZ.** To adjust picks for all **BGEZ** traces, the control file might look like:

/Projects/Montserrat/data/SAC/961102/003443 BGEZ
/Projects/Montserrat/data/SAC/961102/003926 BGEZ
/Projects/Montserrat/data/SAC/961102/004200 BGEZ

For handling multicomponent data:

/Projects/Montserrat/data/SAC/961102/003443 BGEX BGEY BGEZ
/Projects/Montserrat/data/SAC/961102/003926 BGEX BGEY BGEZ
/Projects/Montserrat/data/SAC/961102/004200 BGEX BGEY BGEZ

CCHAR will determine from the configuration of the control file whether single-component correlation or multi-component (with automatic polarity filtering) options are to be employed.

A1.3 Preliminary cross-correlation:

The preliminary cross-correlation step, which involves the full catalogue, is executed to obtain waveform similarity measures for all events. The choice of correlation parameters as well as which station or suite of stations is used for clustering depends very much upon the characteristics of the individual dataset and network. In the two example applications of Sections B and C of this thesis, two very different criteria were used. On the one hand (in the Section C Soultz catalogue of consistent station constellation and known gradual spectrum of waveform types), we chose to estimate preliminary correlation values for the most inconsistent phase, using a rather long window with heavily filtered data; these results were then aggressively decoupled into many small but highly similar families. On the other hand (with the Redoubt catalogue in Section B), we clustered instead using the phase for which we had the greatest number of usable earthquakes within the catalogue (P wave picks at station NCT). Having *a priori* knowledge that the VT and LP volcanic events provide distinct similarity groups, we allowed the "natural" clustering

to dictate group memberships rather than relying upon any rigid cutoff value. Preliminary correlation decisions are also driven by catalogue size and computer system limitations. Storage of several correlation matrices for very large datasets may not be possible.

A1.3.1 Program Use

The correlator may be invoked at command level to obtain an argument list:

corr_sac <cr>

Usage: corr_sac controlfile [P, S] [wlen] [cohpow] [deg, m, none] [istart]

where [P, S] selects phase and [wlen] is length of window in samples

[cohpow] is the exponent of coherency for filtering, squared...

e.g., cohpow=1.0 = cohereency^2

[deg, m, none] SAC header locations are in degrees with z=km (deg);

[istart] integer event number to start (default/blank=0)

controlfile: as discussed above. Path plus filename up to 92 characters long, this limit can be changed via **STRLN** definition in the **corr_sac.h** include file

[P / S]: Which phase do you want to correlate? If you say P, then it will window around the P-pick (header value **a**, minus trace start **b**). If you say S, then it will window about the S (header value **t0**, minus trace start **b**). Note that generally you may want different window lengths for P versus S. This is discussed later.

[wlen]: Number of samples in the correlation window (integer). **Please note** that

although in theory you may choose any windowlength, some lengths seem to make the multitaper calculation unhappy; for instance, it dislikes 36 so you might choose instead 35 or 37.

[cohpow] Coherency exponent that will be squared for adaptive filtering. Cohpow=0.0 switches off the filtering. Cohpow=1.0 filters using coherency squared....etc.

[deg, m, none]: This switch establishes how the geographic information is stored in the trace headers. In the instance of Soultz, everything is in meters. Most earthquake data tend to be stored as lat/lon degrees and depth km. Choose accordingly.

[istart]: OPTIONAL - if you want to start correlating some distance into the matrix, e.g. in case of computer crash. Istart is the index of the first event to begin with (only events > istart will be compared).

A1.3.2 Correlator input

I have a spatial cube of events in the Soultz reservoir (about 100x100x100 m) that contains 696 microearthquakes, based on preliminary hypocenter locations. I isolated these events and want to cross-correlate the P-arrivals at 3-component station 4616. My **control file** looks like this:

/soultz1/data/04061 4616.X 4616.Y 4616.Z
/soultz1/data/04065 4616.X 4616.Y 4616.Z
/soultz1/data/04066 4616.X 4616.Y 4616.Z
/soultz1/data/04069 4616.X 4616.Y 4616.Z

...with 696 lines. I invoke the correlator and direct my output to the file **cube.out**:

corr_sac control/cube_4_8_10.control lag P 64 0.5 xyz > cube.out

A1.3.2.1 Correlator output

Two output files are produced:

Output file (standard output. If not redirected this will go to the screen.)
evfile tracks event name to correlation index

Output file: This is the file which will be passed to the conjugate gradient solver:

0 1 10.809 0.340 0.807 1.0
0 2 -0.277 0.374 0.748 2.350
0 3 5.957 0.257 0.865 0.32
0 4 -7.000 4.081 0.757 0.88
0 5 -1.000 2.690 0.515 5.758
.....

Columns 1 and 2 are the (0:*N*-1) and (1:*N*) indices of the correlated events (event 0 with 1, 0 with 2, ... event 1 with 2, 1 with 3,.....694 with 695).

Column 3 provides the estimated pick lag for each correlated pair. This is in samples, not seconds. Note that sometimes the lag is an integer value with no finer adjustment. These events failed the quality test and were not passed to the subsample correlation routine (**corrsub_fine**), so have been adjusted only to nearest integer sample (**corrsub_coarse**). This test is based on the coarse lag standard deviation combined with maximum correlation coefficient.

Column 4 provides the crosscorrelation standard deviation for this waveform pair. The number is the quadrature sum $\sqrt{a^2 + b^2}$ of the integer and subsample cross-correlation standard deviations. Notice that the large standard deviations correspond to pairs that were not passed to subsample correlation. This standard deviation will be used by the conjugate gradient solver for weighting the constraints.

Column 5 gives the broadband cross-correlation coefficient maximum for the

waveform pair. This value is used, in combination with the integer lag standard deviation, to determine which waveform pairs are passed to the subsample lag estimation routine. This value is also used by the conjugate gradient solver to determine which constraint pairs are rejected *a priori*, before the initial solution and misfit calculations are performed. Further, this value is that upon which waveform similarity clustering is based, in the call to **run_cluster**.

Column 6 reports inter-hypocentral distance, using header values **evla**, **evlo** and **evdp** for preliminary earthquake locations. Numbers reported here are ignored by the conjugate gradient solver and are not used by the correlator unless the distance limit **DMAX** is invoked. If the user so chooses, he may opt to cluster events based on distance rather than correlation coefficient; **run_cluster** will accept these values. Note that proper calculation within corr_sac of the interhypocentral distances depends upon proper selection **[deg, m]** of geographic command-line argument. If hypocenter information is not present in the headers, values of zero will be reported.

To restructure this output into the more familiar matrix format, the system of equations is $\mathbf{A}\vec{\mathbf{x}} = \vec{\mathbf{b}}$, where **A** is the weighted system matrix, $\vec{\mathbf{b}}$ is the data vector of lags and $\vec{\mathbf{x}}$ is the desired vector of pick adjustments,

$$\mathbf{A} = \begin{bmatrix} \frac{1}{0.340} & \frac{-1}{0.340} & 0 & 0 & 0 & \cdots \\ \frac{1}{0.257} & 0 & \frac{-1}{0.257} & 0 & 0 & \cdots \\ \frac{1}{4.081} & 0 & 0 & \frac{-1}{4.081} & 0 & \cdots \\ \frac{1}{2.690} & 0 & 0 & 0 & \frac{-1}{2.690} & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix} \text{ and } \vec{\mathbf{b}} = \begin{bmatrix} 10.809 \\ -0.277 \\ 5.957 \\ -07.00 \\ \cdots \\ \cdots \end{bmatrix}.$$

Evfile: This is the listing that allows you to associate the terse index identifier with a real event.

0 /soultz1/data/03121
1 /soultz1/data/03203
2 /soultz1/data/03214
3 /soultz1/data/03215

A1.3.2.2 Correlator run-time parameters

There are some values which are established in the #include file corr_sac.h, rather than on the command line; these may be modified as needed. The program will need to be re-compiled if this is done. Parameters are described below:

#define MAXSTA 5	maximum number of stations in control file (components).
#define CORRMIN 0.8	correlation minimum for subsample calculation.
#define STDCUT 2.0	minimum st. dev. for fine correlation/
#define NTAPERS 8	number of eigentapers to use for fine st.dev.
#define AVGEIG 2	number of eigentapers phase slope estimate
#define NB 8	number of coarse correlation narrow band masks.
#define ITMAX 1	number of iterative window realignments
#define OFFFRAC 1/4	pre-pick offset fraction of correlation window.
#define DMAX 5000.0	maximum hypocentral separation to consider
#define STRLN 92	maximum string length in filename constructions
#define DOCOHFILT 1	DOCOHFILT=1 to prefilter traces by coherency
#define PFAC 2	zero-padding multiple of trace length for cross-correlation
#define DEGTOKM 111.1	degree to kilometer conversion for spatial considerations
#define ijktoind(i,j,k,idim,jdim) (k)*idim*jdim+(j)*idim+i	indexing definition for 3D array storage
#define KMAX 20	ntaper dimension for multitaper arrays

Of the above definitions, few are likely to need modification; however, those most likely to need changing are as follows:

OFFFRAC: Fraction of correlation window preceding the pick. This gets into the problem of pre-pick contamination. For very quiet pre-P data, having too large a pre-pick offset

will result generally in under-correction of picks. On the other hand, this must be tailored to how bad the picks are. If the window is likely to miss some picks, you may want a longer pre-pick offset.

CORRMIN and STDCUT: Right now if the standard deviation of coarse lag estimate is greater than two samples, we do not pass to fine correlation. Likewise if crosscorrelation maximum is less than 0.8.

Adjust all other control parameters at your own discretion and/or risk.

A1.4 Clustering algorithm

The clustering program, invoked via run_cluster, separates the data either by using the cophenetic correlation for distinguishing very dissimilar families or by using a pre-set correlation cut-off value to terminate joining of events into clusters. If a firm minimum correlation value is desired, the clusterer should be called at command line as:

run_cluster [corr.file] [coeff / dist] [cutoff]

where:

corr_file is the output correlation file from corr_sac

coeff / dist is the choice of clustering on cross-correlation maximum or inter-event distance

cutoff is the correlation minimum or distance maximum for group membership.

If the cophenetic correlation is desired as the clustering criterion, then the cutoff

value should simply be omitted from the command line.

Output from the clusterer will be several files of event indices, each of which represents events belonging to a cluster. The first cluster file will contain all orphan events, if there are events orphaned in the clustering process. These files are named **cluster0000.dat**, **cluster0001.dat**, and so forth. Each cluster file is simply a list of the event indices (0, 1, 2.....) for member events. Associating these with actual events is just a gawk command away. For example, assuming you would like to name the cluster event files **cluster0000.list**:

```
foreach file (`ls cluster*.dat`)
   sort -n $file > temp
   /bin/mv temp $file
   set base = `echo $file | sed `s/\.dat//'`
   foreach value (`cat $file`)
      gawk -v num=$value '{if(num=$1) print $2}' evfile >> $base.list
   end
end
```

If the pre-set clustering cutoff option has been chosen, an additional output file, **cluster.desc**, is generated. This file contains statistical information about the final cluster designations:

clus	N	mean	std	min	max
0	17	0.0490	0.0001	0.0490	0.0490
1	2	0.9650	0.0000	0.9650	0.9650
2	2	0.9490	0.0000	0.9490	0.9490
3	2	0.9450	0.0000	0.9450	0.9450
4	2	0.9380	0.0000	0.9380	0.9380
5	2	0.9300	0.0000	0.9300	0.9300
6	3	0.9383	0.0123	0.9210	0.9480

where:

clus is the cluster number (0000, 0001, 0002....)
N is the number of member events; here, cluster 0 contains 17 orphans.
mean is the mean of the inter-event cross-correlation values in that cluster.
std is the standard deviation of correlation values in the cluster
min is the minimum cross-correlation value of the cluster members

max is the maximum cross-correlation value between any two members.

A good approach is to create a subdirectory for each cluster (i.e. **cluster0000/**), into which its corresponding cluster*.list file is placed. Cross-correlations for all the member events for each station can be thus handled and tracked independently. This is easily automated in a c-shell script.

A1.5 Performing intra-cluster correlations

Once clustering has been completed and event memberships have been assigned, intra-cluster cross-correlations may proceed. Each station will be cross-correlated for all member events for which it has a P and/or S pick. This is very easily handled via c-shell scripting to build and use temporary controlfiles for the correlator, and to perform the correlation repeatedly with different window lengths for each phase. Correlation output files are compared for the suite of windowlengths and the best result (generally some combination of mean correlation value and lowest lag standard deviation) is chosen to pass to the solver. A piece of script that does this is shown below, where **sta1** is the trace of interest, **corrlist** is the list of events in each cluster directory, we are correlating P-waves and the program **$LOCAL/cluster_corrvals** finds the correlation maximum, weighted by lag standard deviation, for the **corr_sac** output file, and reports this in **corrqual.out**.

```
foreach cluster (`ls cluster*.dat`)
cd `echo $cluster | sed `s/\.dat//'`
gawk '{print $0 "sta1.sac"}' $corr.list > test.ctl
foreach winlen (24 32 48 64 72 96 128)
 $CORRDIR/corr_sac test.ctl P $winlen 0.5 m > sta1_$winlen.out
 $LOCAL/cluster_corrvals sta1_$winlen.out
 end
set testval = 0;
foreach winlen (24 32 48 64 72 96 128)
 if -e corrqual.out then
    if(`cat corrqual.out` > $testval) then
   set best = $winlen
   set testval = `cat corrqual.out`
  endif
```

```
 endif
end
/bin/cp sta1_$best.out sta1.out
cd ../
end
```

This loop may be performed for each station to be correlated with the addition of another **foreach....end** cycle in the c-shell script.

A1.6 The conjugate gradient solver:

The conjugate gradient solver has been formulated to operate on the output format from **corr_sac**. The solver (**cgsolve**) is driven by **run_cgsolve**. **run_cgsolve.c** solves the system once and, if the residual rejection option has been selected, it calculates the data misfit vector, sorts it, then iterates through a process of rejecting data outliers and re-solving until the solution reaches an acceptable quality (currently a probability of 0.02 for the reduced degrees of freedom). The resulting output provides a solution which gives pick correction and associated lag standard deviations for each of the waveforms in the data set. This is then read back into the trace headers as header variables **t1** and **t2**, with pick lag standard deviations stored in **user1** and **user2**. in the subsequent algorithm called **addcglags** (discussed later).

A1.6.1 Running the solver:

The solver requires as input the name of the correlation output file, the name of your solution file, an indication of whether to invoke residual rejection for the solution, and an integer number to serve as a seed for the random number generator. Residual rejection is automatically skipped for datasets of fewer than 20 events, although the solver still expects a rej or norej command line argument.

An argument list may be obtained by invoking the program with no arguments:

run_cgsolve <cr>

Usage: run_cgsolve dfile solutionfile [rej/norej] [seed]

where

dfile is the name of the input data file (**corr_sac** output)
solutionfile is the name of the output solution file....
[rej/norej] switch for residual rejection (**rej**), or not (**norej**).
[seed] is an integer seed for the random number generator.

The output file (**solutionfile**) looks like this:

-3.763857 0.064724
-4.982162 0.054454
-0.285896 0.053817
10.525084 0.055334
.....

The two columns represent the pick correction (in samples) and associated standard deviation (in samples). A negative pick lag corresponds to an earlier new pick time. Standard deviations are a reflection of the relative pick standard deviation within the cluster.

If **cgsolve** or its subroutines fail to converge on a solution in a predetermined number of iterations, you will receive associated error messages to your screen. This is usually a question of insufficient constraints with high enough quality (low enough standard deviation and high enough correlation value).

A1.6.2 Solver run-time parameters

The **run_cgsolve** code has an include file of control parameters, **conj.h**. Parameters within **conj.h:**

#define PI 3.1415926	pi
#define STRLN 85	max length of character strings for input files.

#define CORRMIN 0.5	minimum correlation value to use the constraint
#define CGFTOL 1.e-4	convergence tolerance
#define CGITMAX 400	maximum iterations for seeking minimum
#define BCGOLD 0.3819660	Golden ratio
#define BITMAX 1000	maximum iterations for functional minimum search
#define BZEPS 1.e-10	precision tolerance on functional minimum

These four relate to the search for steepest downhill direction

#define MNGOLD 1.618034	Golden section step
#define MNGLIMIT 100	Magnification limit for parabolic step.
#define MNTINY 1.0e-20	Prevents division by zero.
#define MNITMAX 100	Iteration limit on functional minimization.
#define SHFT(a,b,c,d) (a)=(b);(b)=(c);(c)=(d);	sorting / exchange shorthand.
#define REJCOUNT 1/4	fraction of residual vector to shift binary search
#define NREAL 50	Monte Carlo iterations.
#define HUBEREPS 0.1	Functional modification limit near origin.

Of the above parameters, the ones which the user may wish to alter are **STRLN**, **CORRMIN** and **NREAL**, all of which might be increased, depending upon needs. Some efficiency could be gained by reducing the iterations allowed under **CGITMAX, BITMAX** and **MNITMAX**. Changing any parameters in the **conj.h** file will require recompiling to implement.

A1.7 Entering new picks into trace headers

The output from the solver is designed to be fed into the subsequent program (**addcglags**) for stuffing the adjusted picks and their standard deviations into the trace headers. This program currently requires as input a control file of the same form required by **corr_sac**, the phase (P,S,X or Y, see below) to be adjusted and the name of the **run_cgsolve** solution file:

Usage: addcglags [controlfile] [P,S,X,Y] [lagfile]

[controlfile] is the <path> <trace> list of files to alter
[P,S,X,Y] is the pick to adjust:

P,S = correct raw P,S pick
X,Y = correct lag-adjusted P,S picks with stack lag.
[lagfile] is the file containing lags and std. deviations for the adjustments. If P or S is chosen, lagfile must be the same length as controlfile. If X or Y is chosen, lagfile must be a single line.

Note that the original picks remain in the trace headers.

A1.8 Stacking intra-cluster correlations and correlating stacks

Once intra-cluster lag adjustments have been entered into the trace headers, the phases may have stacks produced on the adjusted picks, which may then be cross-correlated for inter-cluster lag adjustments. From an earthquake location standpoint, this is crucial to obtaining proper relative cluster centroid locations. By handling the intra-cluster relative lags first, we have assured the tightest possible constraints and best possible inter-event relationships for the most similar waveforms. Addressing the relative lags between their stacks then finds the relative adjustments needed between clusters and preserves the intra-cluster relationships, which are not altered in the hierarchical stacking process.

A1.8.1 Intra-cluster stacks

The stacking routine is called in much the same way as the correlator, as follows:

Usage: stack_sac controlfile [P, S] [wlen] [stack-filename]

The controlfile will be the same as the controlfile for the initial correlation of a cluster of single-component traces. The **wlen** parameter will determine how long (in samples) the output stacked trace will be. This should be generous enough to accommodate the expected inter-cluster shifts that may occur upon stack correlation, and generally I suggest a few hundred samples. The output file will be named **stack-filename**. It will have in its

header a P or S pick, depending on the command line choice. These will be assigned as **SAC** header variables **a** or **t0**, respectively, which prepares the stacked trace immediately for **corr_sac**.

A1.8.2 Inter-cluster stack correlation

To cross-correlate the cluster stacks, build a controlfile which addresses the stacked traces for all clusters for a particular phase; for instance your control file may be named **P_statAstacks.control** and look like this:

/my/datapath/cluster0001 P_stationA.stack
/my/datapath/cluster0002 P_stationA.stack
/my/datapath/cluster0003 P_stationA.stack

where each **P_stationA.stack** is the stacked trace for P-waves at station A in that cluster. Next, use **corr_sac** to cross-correlate these stacks, and use **run_cgsolve** to find lag adjustments which will shift the stacks to the best possible alignment:

```
>>  corr_sac P_statAstacks.control P 128 0.5 deg > P_statAstacks.out
>>  run_cgsolve P_statAstacks.out P_statAstacks.soln rej
```

Finally, the output stack lags will be added into the original trace headers within each cluster. This is achieved via **addlags**, invoking the **X** or **Y** picktype option. In this case the **lagfile** you offer it will be a one-line file containing the pick lag and standard deviation values corresponding to the output from **corr_sac** operating on the stacks; however, the controlfile will list all member events for the cluster, with traces for station A.

A1.9: Example CCHAR Application

An example of the process is outlined below. Let us assume a hypothetical seismic network of three stations, stationA, stationB and stationC. Our network has accumulated a

catalogue of six earthquakes which for some reason we want to autorepick: event1, event2,.....event6. We will organize our catalogue as follows:

/path-to-data/event1:	**/path-to-data/event2:**	etc.
stationA.sac	**stationA.sac**	
stationB.sac	**stationB.sac**	
stationC.sac	**statoinC.sac**	

Our control file, named **stationA.control**, for the initial correlation on stationA will then look like:

/path-to-data/event1 stationA.sac
/path-to-data/event2 stationA.sac
/path-to-data/event3 stationA.sac
/path-to-data/event4 stationA.sac
/path-to-data/event5 stationA.sac
/path-to-data/event6 stationA.sac

We will correlate on the P-waves for stationA, as we have decided this is the phase we wish to use to determine catalogue clustering divisions.

corr_sac stationA.control P 256 1.0 deg > stationA.corrout

This produces an output file stationA.corrout which looks like:

0	1	-5.234	0.013	0.997	12.34
0	2	-4.013	0.003	0.921	6.113
0	3	-3.133	1.344	0.753	22.51
0	4	-2.000	2.377	0.512	15.11
0	5	-0.700	1.103	0.857	9.033
1	2	1.001	0.035	0.953	7.001
1	3	2.379	..... and so forth.		

Also generated was a file, **evfile**, which tracks events and event indices:

0 /path-to-data/event1
1 /path-to-data/event2
2 /path-to-data/event3
etc...

We pass **stationA.corrout** to **run_cluster**:

run_cluster stationA.corrout coeff

RUN_CLUSTER v0.0.1 : Dendrogram-based data clustering algorithm developed for use on seismic waveform similarity matrix output from corr_sac. Modified from the Sandia National Labs MATLAB-based MatSeis software. C. Rowe, 14 March 2000

Counting similarity pairs:
6 events were correlated
vector array is 15 entries long
Begin memory allocation.
memory allocated, begin scanning correlation data.
Begin clustering....
Clustering complete.
Dividing clusters into output lists.
finding maximum drop in cophenetic correlation.
stopping clustering at index 3
this leaves us 2 groups

The two output files, **cluster0000** and **cluster0001** appear as follows:

cluster0000:	**cluster0001:**
0	**3**
1	**4**
2	
5	

These are the event indices (from 0 to 5) of the events which were correlated. We like to keep our clusters separated into their own directories, so we build our cluster directories and create a correlation controlfile within each:

```
foreach cluster (`ls cluster*.dat`)
set base = `echo $cluster | sed `s/\.dat//'`
mkdir $base
sort -n $cluster > ttt
/bin/mv ttt $cluster
foreach index (`cat $cluster`)
gawk -v num=index '{if($1==num) print $2}' evfile >> $base/corr.control
end
```

We now have two directories, **cluster0000** and **cluster0001**, each of which contains a file called **corr.control**. This file lists the member events for that cluster. We may then use **cluster0000/corr.control** to generate control files and cross-correlate each of our stations for events belonging to **cluster0000**, and likewise for **cluster0001**. After we have correlated and solved (**run_cgsolve**) for all four stations in each directory, perhaps both for P and S, we will have output lag adjustment files in each of the two directories named, let us say, **PstationA.soln**, **SstationB.soln** and so forth.

We enter these lags into their proper traces (and places) with **addcglags**, then create a stack for each phase,

stack_sac PstationA.control P 256 PstationA.stack

yielding files such as:

cluster0000/PstationA.stack **cluster0000/PstationB.stack,**
cluster0000/SstationC.stack **cluster0000/SstationB.stack**
cluster0001/PstationA.stack **cluster0001/SstationA.stack**

The file **cluster0000/PstationA.stack** will be a 512-sample trace that is the stack of all four stationA P windows in **cluster0000** after they have been aligned on their intra-cluster pick lags. It has a P-pick which corresponds to the mean adjusted pick for those four waveforms.

To obtain the inter-cluster pick lag adjustment for station A P-picks, we create a control file (named, say, **PstationAstack.control**):

/path/to/cluster0000 PstationA.stack
/path/to/cluster0001 PstationA.stack

then

corr_sac PstationAstack.control P 96 0.5 deg > PstationAstack.corrout

run_cgsolve PstationAstack.corrout PstationAstack.corrsoln rej

This gives us a solution file of lags which would look like:

-1.3572 0.11332
1.3572 0.11332

since there are only two stack traces being correlated. We may then create a lag adjustment file for each stationA P pick involved in the stacks:

cluster0000/PstationAstack.soln: **cluster0001/PstationAstack.soln**
-1.3572 0.11332 **1.3572 0.11332**

Then introduce the stack-lag adjusted P picks into the member events:

```
foreach cluster (cluster0000 cluster0001)
cd $cluster
addlags PstationA.control X PstationAstack.soln
cd ..
end
```

As stated at the beginning of this appendix, the glue which binds **CCHAR** together is the c-shell scripts. These sequentially invoke the individual C programs and manage the data and file structures. The programs are individually useful for many applications; users may wish to determine their own sequence of program steps (for instance, skip the clustering for events which are similar enough....use a fixed correlation window for all phases rather than scripting it adaptively, etc.). In a single, large package this kind of flexibility would be a problem, and it is understood that not all datasets are created equal. C-shell scripts are included on the enclosed CD which may serve as useful skeletons for user needs, with appropriate path and station name modifications.

CCHAR is very much a work in progress. Upgrades, improvements, enhancements or fixes will be ongoing. Every effort will be made to post maintenance notices, new versions and fixes to the **CCHAR** websites (**http://www.ees.nmt.edu/Geop/cchar.html** and **http://www.geology.wisc.edu/~char/cchar.html**), along with current contact infor-

mation for problems and questions.

To contact the author for questions of problems, try **char@dutchman.nmt.edu** or **char@geology.wisc.edu**. If neither of these is valid (after some years) I should be findable through the AGU or SSA online member directory listings.

Part B: Relocation of Seismicity Associated with the 1989-1990 Eruption of Redoubt Volcano, Alaska: Part II. Application to Redoubt Eruption Seismicity

ABSTRACT

A new algorithm for adjusting phase picks in digital waveform catalogues is applied to event-detected seismicity associated with the 1989-1990 eruption of Redoubt Volcano, Cook Inlet, Alaska. Hypocenter locations using preliminary phase picks describe a diffuse cloud of seismicity containing few if any well-defined structures. Location artifacts resulting from systematic preliminary picking biases produce misleading temporal trends, and source region dimensions of ~0.7 km^3 for volcanic long-period (LP) earthquakes, as well as apparent pipe-like distributions among volcanotectonic events. After we apply our automatic repicking algorithm, long-period events collapse into a much smaller volume (~0.02 km^3), which may represent a stationary volumetric source. Volcanotectonic earthquakes exhibiting sufficient signal quality relocate into tight clusters of seismicity, illuminating a disjointed zone of separate seismic source regions within the previously continuous "pipe," as well as joint or fracture features within the deeper portions of the volcano.

Cross-correlation of LP events reveals that highly repeatable short-period onsets resemble small volcanotectonic events. LP events may thus be triggered by brittle failure adjacent to the resonating volume, in much the same manner as so-called "hybrid" events. Inconsistent correlation lags between high frequency onsets and the volumetric resonance represents variable time and/or distance between repeatable short-period triggers and sub-

sequent volumetric excitation.

INTRODUCTION

Among the more challenging real-time problems that seismologists address is the monitoring of volcanic activity. Eruption forecasting depends significantly upon interpretation of volcanic seismic signals. The quality of the seismic data and derived parameters may ultimately affect the ability of the observers to determine the level of concern that is appropriate during an eruption crisis. Chief among these concerns are the robust discrimination of different families of volcanogenic seismic sources, temporal variation in seismicity patterns and location of events within the problematic and noisy environment often found in such situations. These tasks are further complicated by the often sparse station coverage in volcanic seismic networks, occasionally high mortality rate of the nearest recording stations to the volcanic edifice (e.g., Power et al., 1994), as well as the extremely complex three-dimensional structure of most volcanoes, which lends an additional challenge to the problem of precise earthquake location.

Hypocenter relocation studies have been undertaken in volcanic settings using a variety of approaches. Systematic location biases and distortions arise from inadequate modelling of the complex velocity structure of a volcano; much attention has been paid to improving velocity modelling (e.g. Scarpa, 1996). Refinements to velocity models have included: improved 1-D models (e.g., Rowe, 1988; Lahr et al., 1994; Dawson et al., 1996) for use with single event location (SEL); methods using joint hypocenter determination (JHD) (e.g. Pujol and Aster, 1990) for reduction of systematic traveltime residuals to accommodate receiver-local velocity heterogeneity; forward two- and three-dimensional raytrace algorithms (e.g., Pearson, 1977; Luetgert, 1984); and three dimensional seismic

tomography (e.g., Lorenzen, 1994; Benz et al., 1996; Mori et al., 1996). In addition to sytematic hypocenter mislocation arising from velocity model problems, there is a random component to hypocenter estimates that arises directly from phase picking inconsistencies. Correcting the velocity models through the use of refraction shots, JHD or tomography may adjust the general centroid of hypocenter locations to be approximately correct; however, picking inconsistencies continue to compound the ability to see interpretable structures or trends within the cloud of hypocenters.

Lahr et al. (1994) demonstrated in a study of location errors for a swarm of LP events at the onset of the eruptive activity at Redoubt Volcano, Alaska (discussed in more detail below), that a ~1.5 km "pipe" of hypocenters was statistically indistinguishable from a single point source when picking errors were accounted for. Jones and Stewart (1997) demonstrated spectacular improvement in delineation of ring faults at Rabaul Caldera, Papua New Guinea through application of a technique called "collapsing," which addresses random hypocenter location scatter resulting from picking inconsistency. In this technique, hypocenters are assumed to cluster onto lines or planes, and neighboring events are permitted to move towards one another onto such structures if such perturbation does not significantly increase the location misfit. Application of the collapsing technique to seismicity at Rabaul resulted in significantly improved imaging of ring faults around two nested caldera structures (Jones and Stewart, 1997). An adaptation of this technique was demonstrated by Fehler et al. (2000), in which they incorporating the collapsing equations into the JHD process, so it operates on raw traveltime data.

Since the advent of routine digital seismic recording it has become possible to correct for picking inconsistencies by applying correlation methods to compare the wave-

forms and correct picking inconsistencies among similar events, thus reducing random scatter in the hypocenter cloud. Fremont and Malone (1987) used waveform cross-correlation techniques to obtain precise relative relocations for events at Mount St. Helens, and Gillard et al. (1996) investigated magmagenic seismicity trends at Kilauea with a similar technique. Jones et al. (in press) report on a preliminary investigation of precise relocation of events at Mount Pinatubo, using the Matlab-based Xadjust software (Dodge et al., 1995), which goes beyond relative traveltime differential methods to adjust the phase picks themselves after cross-correlating traces. This approach has the advantage of providing adjusted picks which may subsequently be introduced into other seismological applications.

In a similar approach to Dodge et al. (1995), we have developed an adaptive, automatic, correlation-based phase repicking algorithm for use on digital seismic waveform catalogues, which identifies similar event families and adjusts phase picks for consistency within and among families (Aster and Rowe, 2000, Rowe et al., 2000a). We believe our algorithm provides an advancement beyond the Dodge et al. (1995) technique and others that have gone before it, as we not only correct for intra-cluster relative locations, but also inter-cluster centroid relationships, without relying upon preliminary (and possibly biased) catalogue parameters. The technique does not require the Matlab software and is optimized to handle very large systems efficiently, both in terms of CPU and computer memory requirements. Computational and algorithmic details of the technique are presented in Aster and Rowe (2000) and Rowe et al. (2000a). In this study we demonstrate our method by applying it to the digital waveform catalogue of volcanic seismicity associated with the 1989-1990 eruption of Redoubt Volcano, Alaska.

DATA

In September, 1989, background seismicity at Redoubt Volcano, Cook Inlet, Alaska (Figure 1), began to exhibit small LP events, with minor swarms in November and December. VT events were observed beginning in November, but overall seismicity rates were still very low (Power et al., 1994). On December 13, 1989, an intense swarm of LP events began (Power et al., 1994; Lahr et al., 1994; Stephens et al., 1994), which merged into continuous tremor by December 14, leading later that day to the first large explosion of the eruption sequence. Several subsequent episodes of relative quiescence, seismically energetic dome-growth and subsequent dome-destroying explosions, as well as debris avalanches, pyroclastic flows and volcanic tremor, were recorded by the Alaska Volcano Observatory (AVO) and Alaska Earthquake Information Center (AEIC) systems. Eruption and post-eruption seismicity continued at some elevated level until approximately June, 1990, at which time activity returned to background levels (Power et al., 1994).

Data acquisition for the Redoubt eruption was carried out in both analog (helicorder) and digital modes. The combined AEIC and AVO network provided digital recording (Sonafrank et al., 1991) for Cook Inlet and Redoubt. Both of these systems operated in event-detection mode. AEIC also implemented continuous digital recording (e.g. Rowe and Davies, 1990; Tytgat et al., 1992) prior to the eruption onset.

Figure 2 shows the locations of AEIC/AVO seismometers which were used in recording the Redoubt seismicity during the course of the 1989/1990 eruption sequence. Network configuration changed somewhat during this time, as new stations were added and/or stations were destroyed or disabled by the eruption or related lightning strikes.

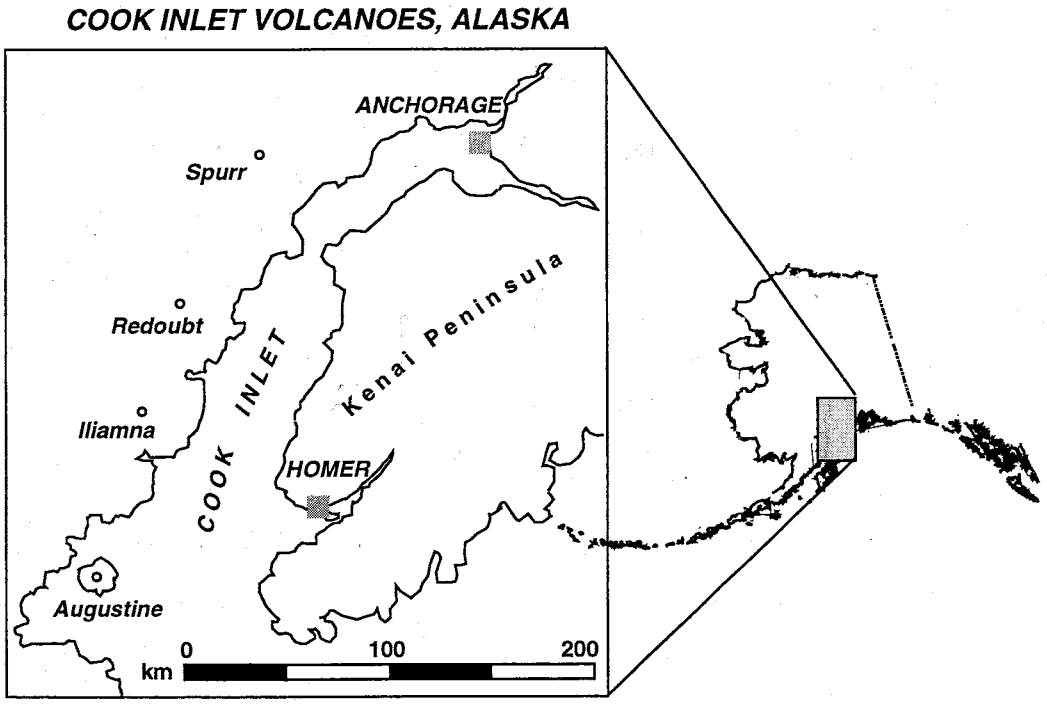


Figure 1: Map of Cook Inlet, Alaska, showing four volcanoes: Redoubt, Spurr, Iliamna and Augustine Island. Also indicated are the locations of the cities of Anchorage and Homer.

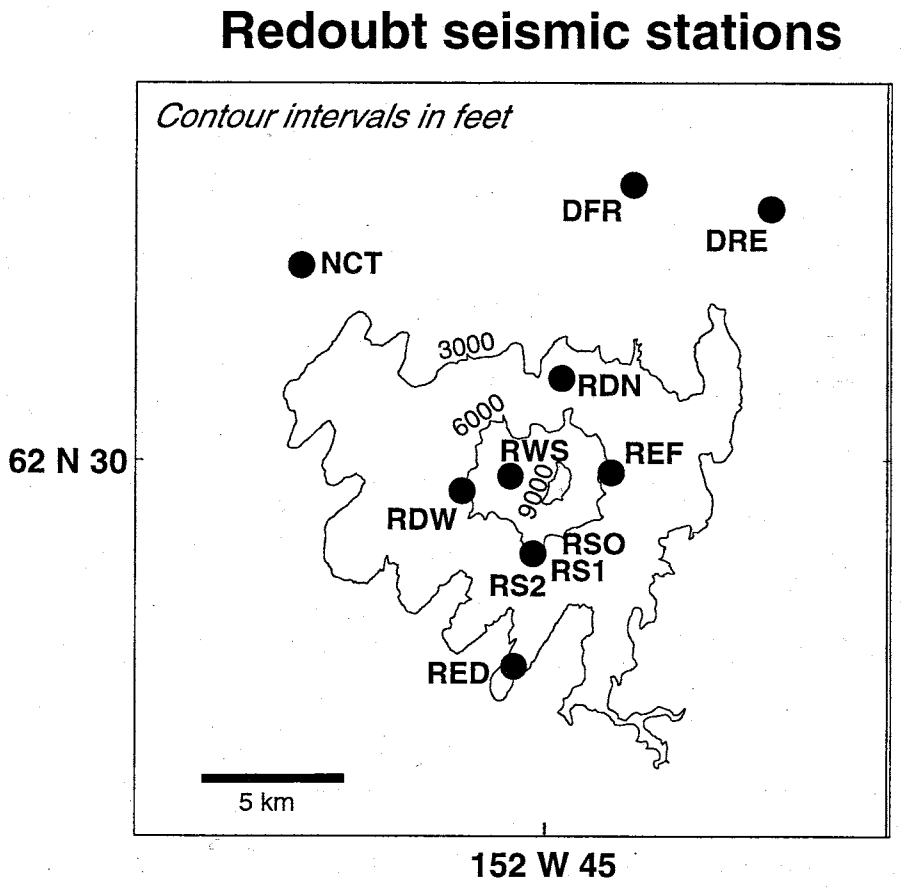


Figure 2: Map of Redoubt Volcano, Alaska and AEIC/AVO seismometers which operated during the course of the 1989/1990 eruption.

Signal saturation (clipping) for most stations which were also near enough to record an appreciable number of earthquakes is a problem. To reduce waveform clipping, some stations had A1 type voltage-controlled oscillators (VCOs), which allowed gain-ranging (Rogers et al., 1980). When signals reach an amplitude of 46 dB, the station automatically drops in gain by a factor of 20 dB (Rogers, 1980; Power et al., 1994). Although this does permit determination of maximum amplitude (and hence magnitude) for the largest events, it results in signals which are difficult to reconstruct. The distortion, as well as the compromised signal envelope, make digital signal cross-correlation problematic for many events recorded at such stations. Figure 3 shows five traces from station RDN, illustrating the characteristic gain-ranging distortions. Note that the gain-ranging in Figure 3 commences at different points following event onset, depending upon the size of the earthquake. In the uppermost panel of Figure 3 we indicate onset of gain-ranging (1), indicator spike (2) which is used to confirm the station has switched into gain-range mode, and the exit from gain-range status (3) when signal strength drops below the required threshold.

Despite these contaminating nonlinearities, we had success cross-correlating the waveforms when we applied the adaptive, coherency-based filtering and used a sufficiently long correlation window so that distortions comprise only a small part of the time series in question. The scaled amplitude differences between coherent parts of the signal are immaterial since signals are normalized before cross-correlation. Stations RDN, NCT, DFR and RDT (Figure 2) were all initially equipped with automatic gain-ranging; the feature was disabled at station RDN on January 9, 1990 because it complicated real-time estimates of seismic amplitude levels during the eruptions (Power et al., 1994). Figure 4 illustrates waveform alignments for station RDN for the December 13-14 LP swarm (dis-

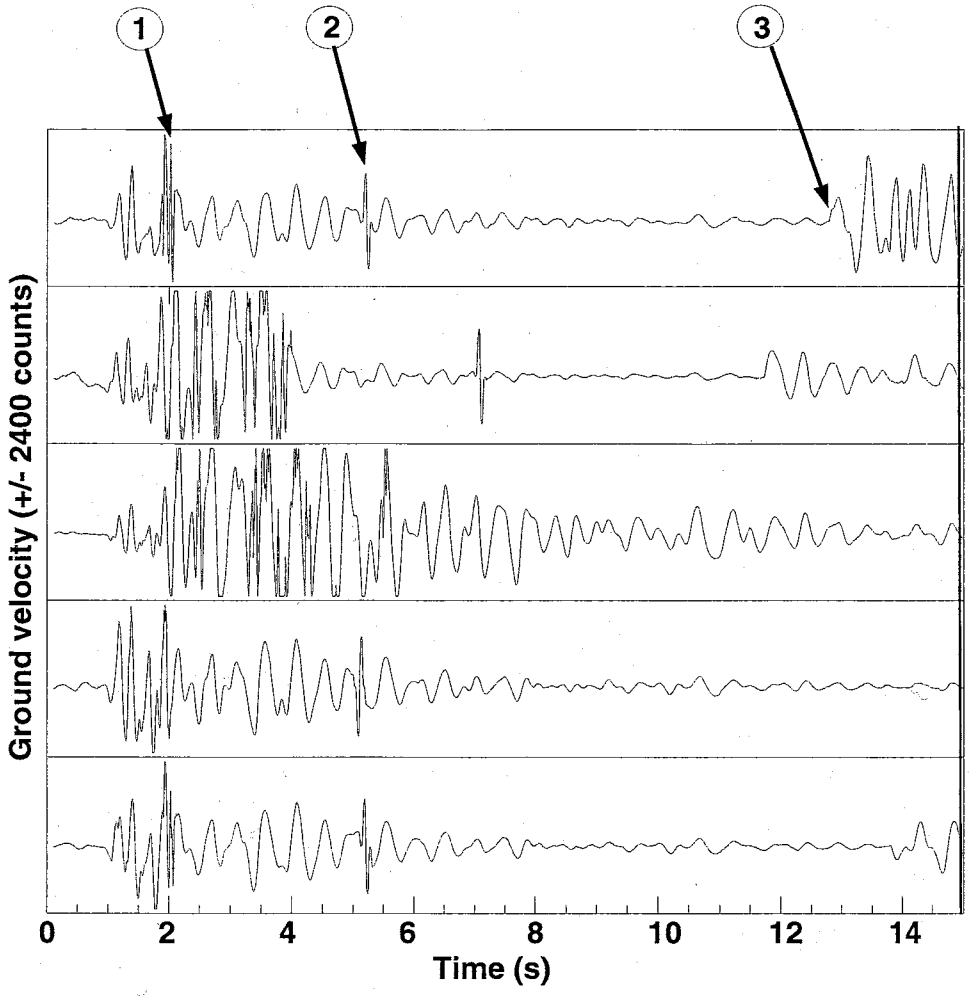


Figure 3: Typical seismic signals exhibiting A1-VCO gain-ranging characteristics. These five seismograms, all recorded at station RDN, illustrate the signal distortions that occur when gain-ranging is enabled at a seismic station. Indicated on the uppermost trace are 1) signal saturation and 20 dB drop, 2) subsequent indicator spike to confirm entry into gain-ranged mode and 3) exit from gain-range status with declining signal amplitude. These nonlinear signal distortions, which occur at different times depending on the strength of the ground motion, may complicate cross-correlation efforts.

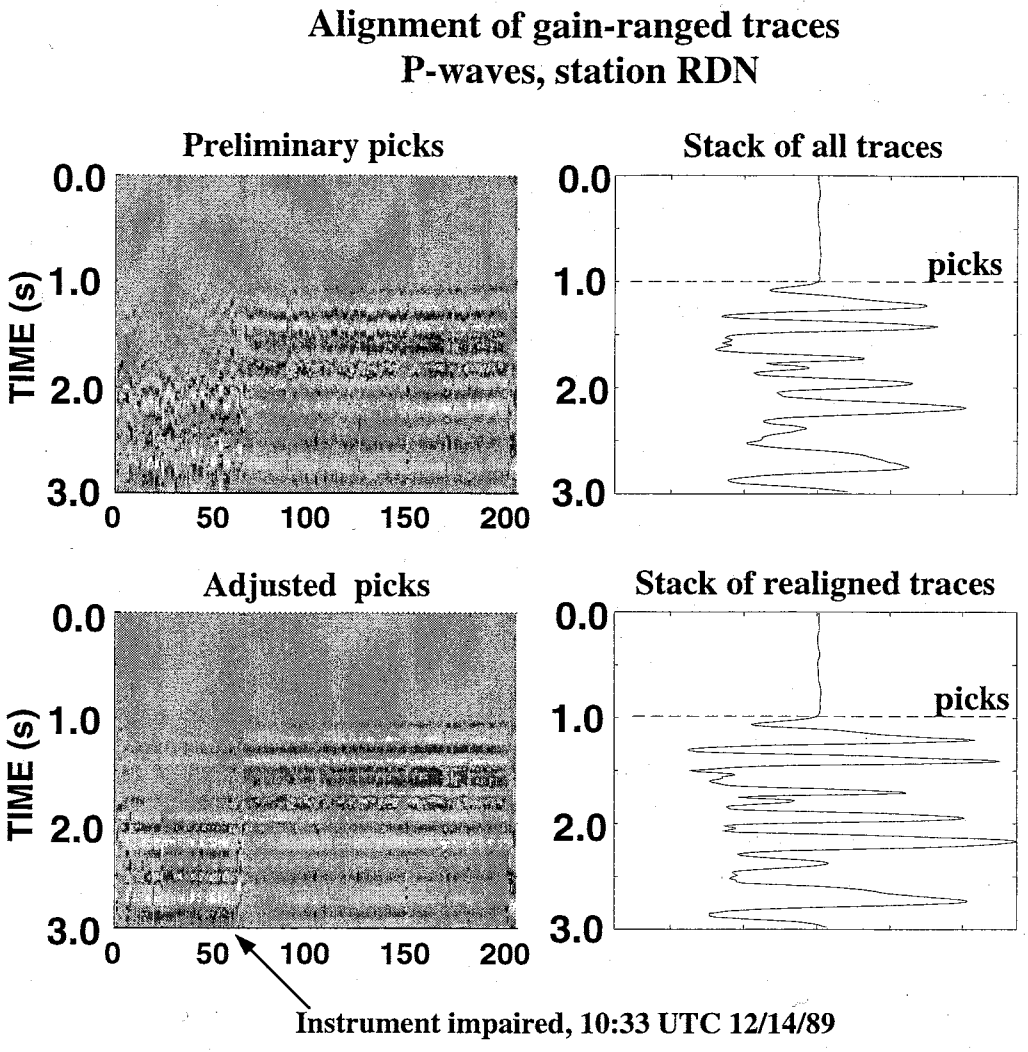


Figure 4: Waveform alignment plot for station RDN, showing P-wave alignments for events of the December 13-14 LP swarm. The sudden change in signal character clearly indicated at event number 63 is the result of damage resulting from early eruptive activity. The signal nonlinearities resulting from the gain-range feature on this instrument constitute an incoherent and minor portion of the waveform spectra if sufficiently long correlation windows are used. Application of the coherency-weighted, adaptive filter permits successful cross-correlation and improvement in pick consistency.

cussed in more detail below), illustrating the ability to cross-correlate these signals despite variable onset of gain-ranging among earthquakes. Note the sudden change of signal character in Figure 4, easily identifiable at event number 63, corresponding to 10:33 UTC on December 14. Inquiries to AVO revealed that this station was damaged in the initial vent-clearing eruption at 10:30 UTC on December 14, 1989 such that its frequency response was altered (J. Power, pers. comm; Power et al., 1994). Although impaired, the station continued to record events and was clearly still useful for hypocenter location. Our cross-correlation method succeeded in providing consistent waveform alignments despite these irregularities. Additional network changes were noted via waveform alignment plots for station RSO, which experienced a clear (and permanent) polarity reversal during June, 1990 (Figure 7). Station history records documented a site visit on 29 June; the first reverse-polarity trace occurs on 30 June. Observations of this nature show the robustness of our adaptive filtering and clustering, while incidentally suggesting possible retroactive or near-real-time network diagnostic applications for such cross-correlation and similarity clustering techniques.

We have examined triggered digital waveform data for Redoubt Volcano for the years 1989-1994, inclusive. The data include all available event-detected digital seismic data for the eruption, as well as some missed events which were restored from the AEIC continuous digital archive (e.g., Rowe and Davies, 1980). Preliminary hypocenter locations and phase arrivals (Power et al., 1993) were obtained in HYPOELLIPSE pickfile format (Lahr, 1989), as well as the working velocity model for routine observatory hypocenter location.

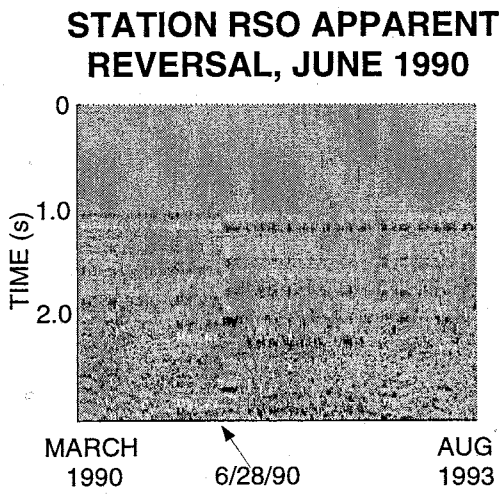


Figure 5: Waveform alignment plot for station RSO, for events from one similarity cluster with membership from March 1990 through August 1993. Note the sudden reversal of peaks and troughs indicated by the arrow; this (permanent) polarity reversal coincides with updated station parameters found in the station history log and indicates a site visit during late June 1990. Such observations may provide useful retroactive network diagnostic information

PROCEDURE

One of our goals is to demonstrate the applicability of this technique in routine observatory operations, so instead of completing the catalogue for homogeneity prior to cross-correlation, we have only addressed waveforms and their P and/or S phases for which preliminary picks were available. The resulting improvement in hypocenters may therefore not necessarily be optimal, but we may demonstrate the degree of improvement which might be realized on an ongoing basis during an eruption crisis. To assess our ability to improve resolution with methods routinely employed in observatory operations, both preliminary (using original picks) and revised locations (using automatic repicks) were calculated using HYPOELLIPSE (Lahr, 1989) with control files and velocity models obtained from AVO. Because of variable network configuration throughout the course of the eruption, the catalogue is heterogeneous in terms of hypocenter location error parameters, and the events included in the correlation and solution for pick lag adjustments will necessarily change on a station-by-station basis. We proceeded by first addressing the signal processing and lag correction aspect of the problem, adjusting the inconsistent picks among similarity clusters on a station-by-station basis. A comparison of repicked hypocenters with the preliminary hypocenter locations will demonstrate the benefits gained by the automatic repicking and clustering algorithm, irrespective of systematic location improvements which might be gained by improving the seismic velocity model.

Grooming the catalogue

Implementation of the repicking algorithms inevitably requires a certain degree of catalogue grooming. The first task for this study was the isolation of Redoubt-related seismicity from the larger dataset. This was done on a geographic basis to eliminate seismicity

from the other nearby volcanoes in the catalogue, particularly Mount Spurr (Figure 1), whose 1992 eruption is also included in this time period. Events were selected whose preliminary hypocenters were located between 60.25 and 60.75 N, 152.3 and 152.8 W. Preliminary hypocenters were obtained using HYPOELLIPSE (Lahr, 1989) with the 1-D velocity model shown in Figure 6. This velocity model was derived with the help of an active seismic experiment conducted in July, 1991 using IRIS/PASSCAL instruments to augment the AVO network, and four man-made explosions (Dawson et al., 1996). Figure 7 shows preliminary hypocenter locations for the dataset (Power et al.,1993). Only events located using five or more phases are shown. The resulting subset of seismicity comprised 4796 events, October, 1989 through December, 1994. All waveforms for the selected events were converted into SAC format, and preliminary pick and hypocenter information were copied from the HYPOELLIPSE phase archive files into the SAC trace headers.

Clustering and Correlation

Events were clustered based on similarities derived from an extended correlation window (1.28 s) for station NCT (Figure 2), which provided phase arrival times for the greatest number of events. Initial clustering of the resulting, 4106-event catalogue using the cophenetic correlation parameter resulted in ten large families of events with sizes ranging from 187 to 820 members.

After the initial large clusters were identified, member events were cross-correlated with optimal correlation windows for each station and each phase. Because of catalogue inhomogeneity, any given station may have significantly fewer members than the initial, NCT-based division. For instance, in Cluster 9 we found 820 P-wave picks at station NCT; these were cross-correlated and their picks adjusted with respect to the entire

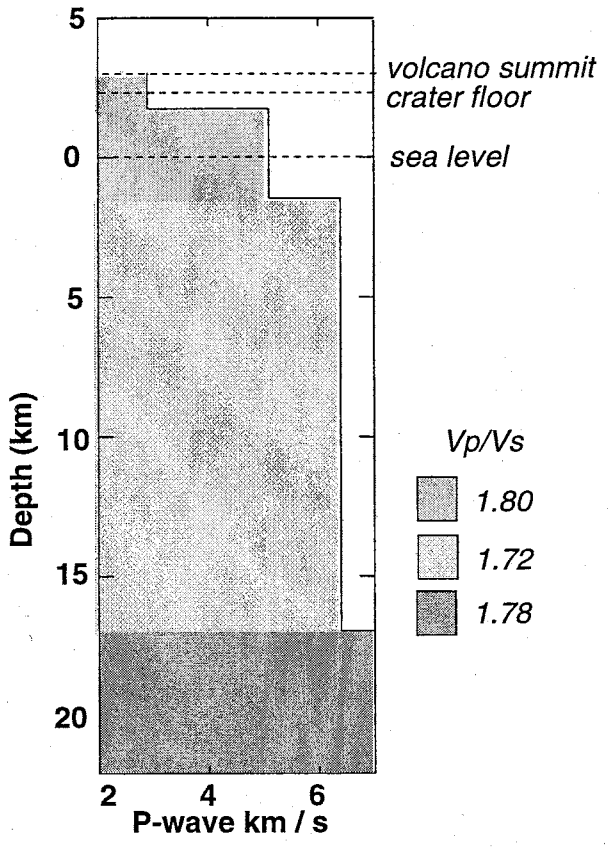


Figure 6: Redoubt one-dimensional velocity model used for event location. Note that "zero" depth is chosen as sea level. Earthquakes located within the edifice of the volcano may have positive elevation. The summit of the volcano is modelled at +3 km, whereas the crater floor resides at an elevation of 2.3 km above sea level.

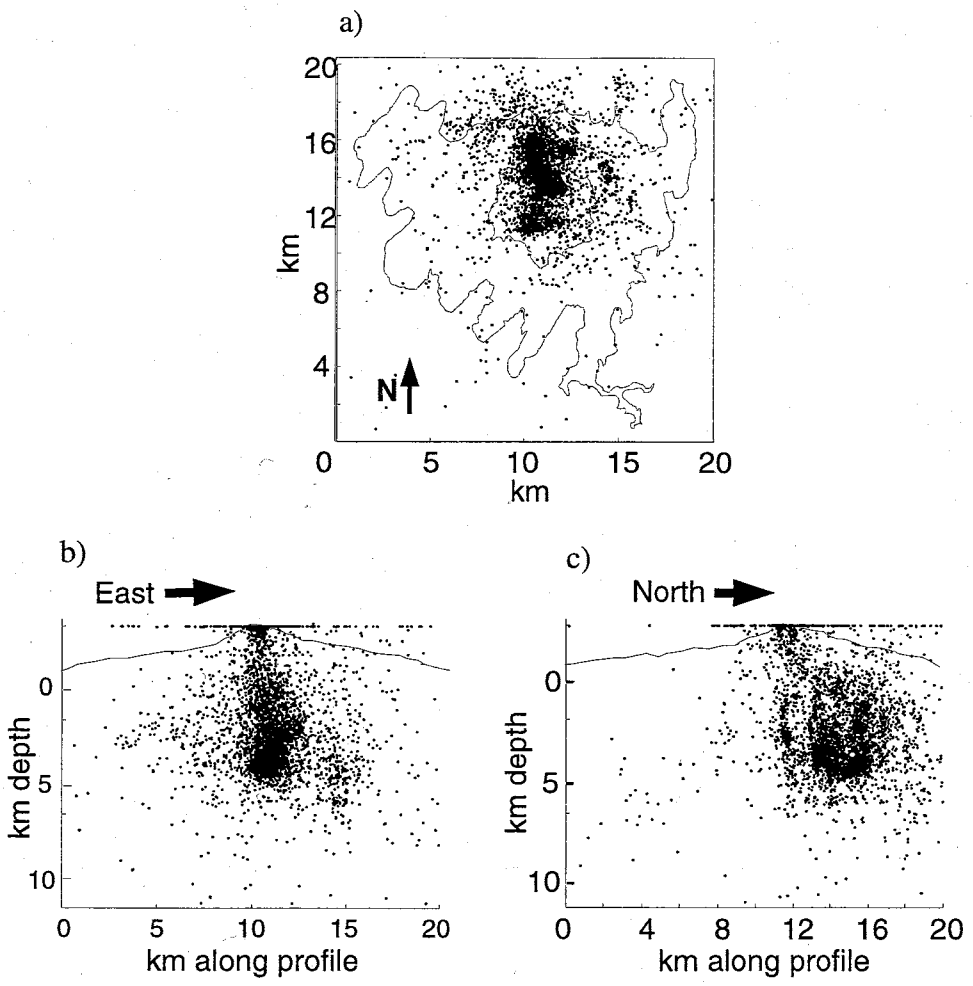


Figure 7: Preliminary locations for the Redoubt seismic catalogue. over 4700 earthquakes are plotted; these were selected from the full catalogue based upon a minimum of five phase picks. No *a priori* selection based on hypocenter RMS or error ellipsoid was used, as we expect these values to change upon relocation. a) Map view of seismicity also showing contours of the volcanic edifice b) East-west cross-section. c) North-south cross-section. Hypocenters aligned at -3 km level represent unstable solutions which were fixed by HYPOELLIPSE to the top of the velocity model.

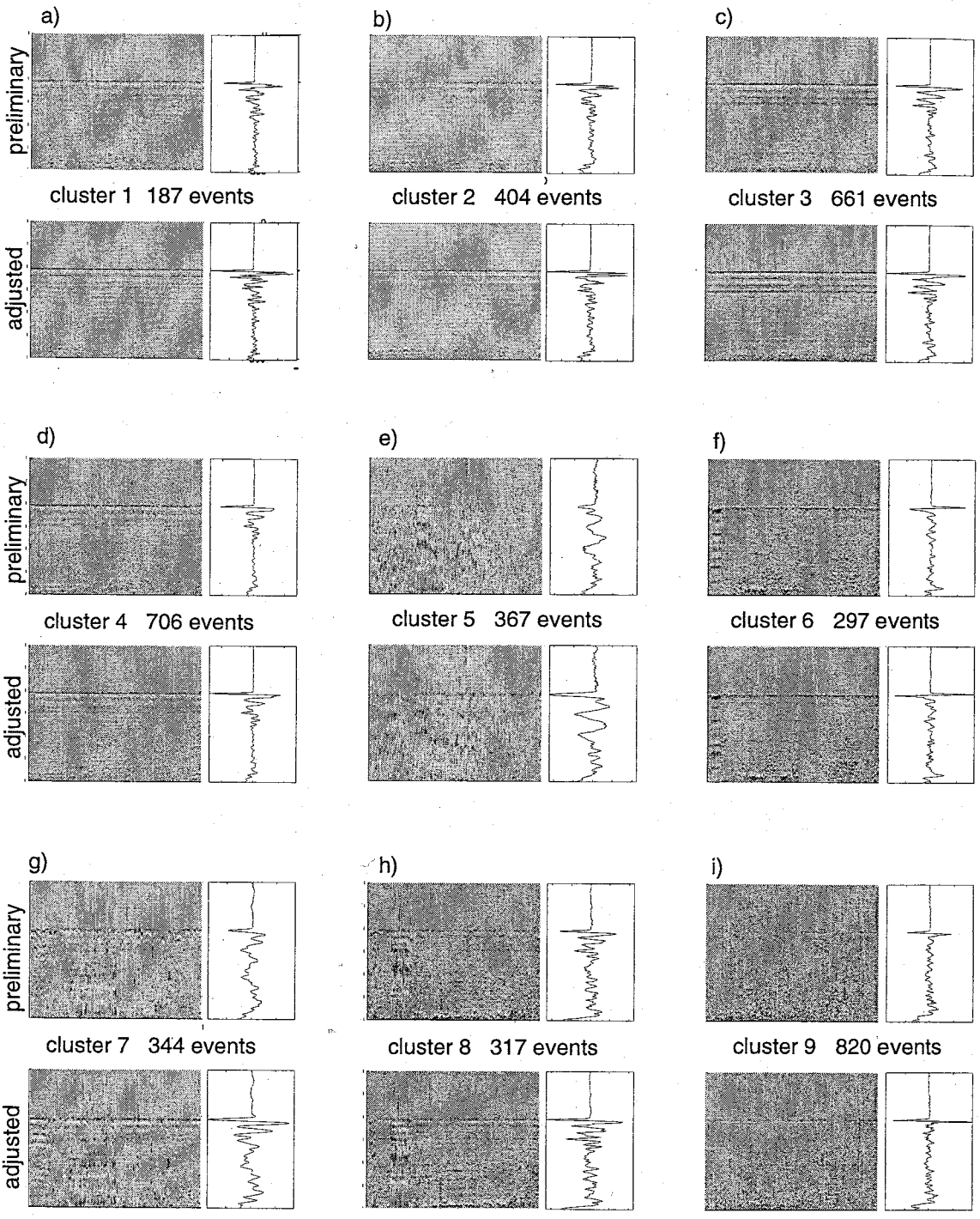
cluster membership. Among these earthquakes, station RDN (see Figure 2) was picked by the analyst for only 294 of the earthquakes; hence our P-wave cross-correlation at station RDN for this cluster will compare only 294 of the member events. Only 554 events had S-wave picks at station NCT, and RDN S-wave picks numbered only 249.

We find that although P-wave arrivals are generally well-picked with reasonable consistency, S-wave picking inconsistencies contribute significantly to the apparent scatter of sources. Figures 8a-8i illustrate waveform alignment plots for station NCT for nine of the ten identified clusters; Cluster 10, corresponding to the December 13-14 LP swarm, will be discussed independently. Upper panels in each figure show traces aligned on preliminary picks (shown as a surface rather than individual traces for interpretability at this scale) and associated stacked waveform; lower panels show surfaces of traces aligned on the automatic repicks, with the resulting stacked waveform. Total dataset P-wave pick adjustments for this station spanned 0.29 s with standard deviation of 0.023 s. The S-wave adjustments spanned a range of 1.22 s, with standard deviation of 0.11 s.

RESULTS

The Redoubt eruption has been roughly divided into six phases (e.g., Power et al., 1994). Each phase is characterized by varying seismic behavior corresponding to the eruptive behavior of the volcano. With the exception of the energetic LP event swarm prior to the first large tephra explosion of 14 December, 1989, the event similarity clusters, which break down roughly into VT or LP signals, all include earthquakes from multiple eruption phases, spanning the ~1 year duration of the volcanic activity. We present our relocation results in the context of these six eruption phases:

Figure 8: Waveform alignment plots and associated waveform stacks for P-waves at station NCT for nine of the ten clusters (Cluster 10 is treated in detail below). We chose this station for our clustering divisions, as it had the greatest number of earthquake phases (4106) picked. Details of cluster membership characteristics may be seen in the stacked waveforms (right hand panels in a-i) for the adjusted alignments (lower panels).


a)
b)
c)
preliminary
cluster 1 187 events
cluster 2 404 events
cluster 3 661 events
adjusted
d)
e)
f)
preliminary
cluster 4 706 events
cluster 5 367 events
cluster 6 297 events
adjusted
g)
h)
i)
preliminary
cluster 7 344 events
cluster 8 317 events
cluster 9 820 events
adjusted

Phase 1: Precursory phase (1 Oct. - 14 Dec., 1989)

The early activity within the precursory phase was subtle (Power et al., 1994), with a few VT earthquakes and low-level occasions of tremor which were only identified retroactively. The later development was dramatic: a vigorous and accelerating swarm of LP events merged eventually into high-amplitude tremor, culminating in the first explosive eruption on 14 December.

LP swarm of 13-14 December, 1989

At approximately 10:30 a.m. (AST) a vigorous swarm of LP events commenced, which continued until the first explosive eruption at ~9:47 a.m. on 14 December (Power et al., 1994). Over 4000 earthquakes were identified within the swarm, before continuous tremor rendered the signals impossible to separate (Chouet et al., 1994). We relocated 187 LP events within the 219 earthquakes comprising our Phase 1 subset. Figure 9 show map-view and a cross-section view of preliminary (left) and relocated (right) hypocenters for these events. Preliminary locations delineating a volume approximately 750 m in diameter extending from 0.6 to 2.7 km beneath the volcano crater floor. Lahr et al. (1994) examined the distribution of LP hypocenters in terms of picking error. They found the greatest scatter among the smallest events, whose locations outline a volume of ~2 km^3, whereas larger events occupy a volume of less than 0.07 km^3. Lahr et al. (1994) further noted that waveform and spectral similarity suggest these signals arise from a very restricted source region. Synthetic locations calculated by perturbing picks within the observed standard errors indicated that the extended LP cluster could not be distinguished statistically from a point source (Lahr et. al., 1994). Our relocated hypocenters (Figures 9c,d) show a reduction of the LP volume to a much smaller source region of approximately 0.02 km^3.

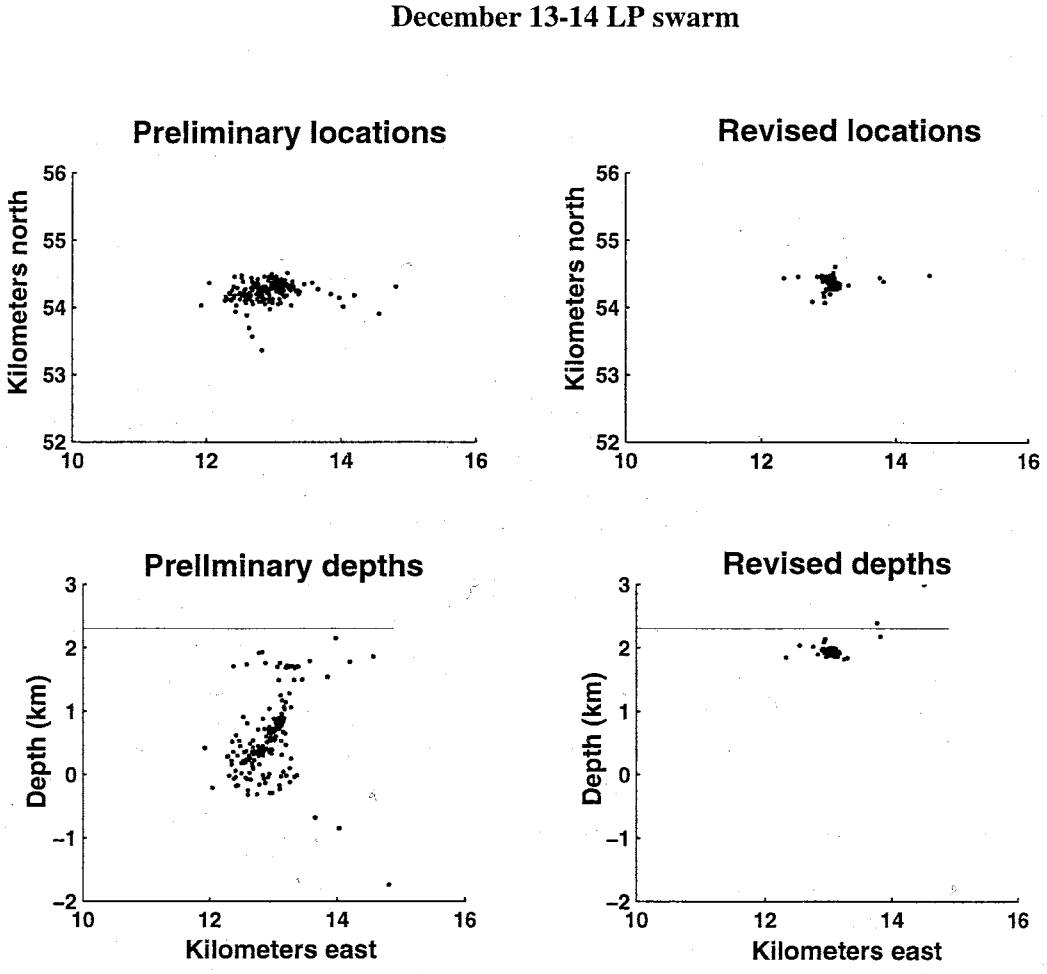


Figure 9: Preliminary and revised locations for the 202 Cluster 10 events, dominated by the December 13-14 LP swarm. Preliminary locations (left) delineate a volume of roughly 0.75 km^3 and the cross-section (lower panel) suggests a "pipe" or "conduit" filled with long-period sources. Relocations of these events demonstrate the source is a much smaller region concentrated into a volume of ~0.02 km^3. Horizontal line in the cross-sectional views indicates the location of the crater floor.

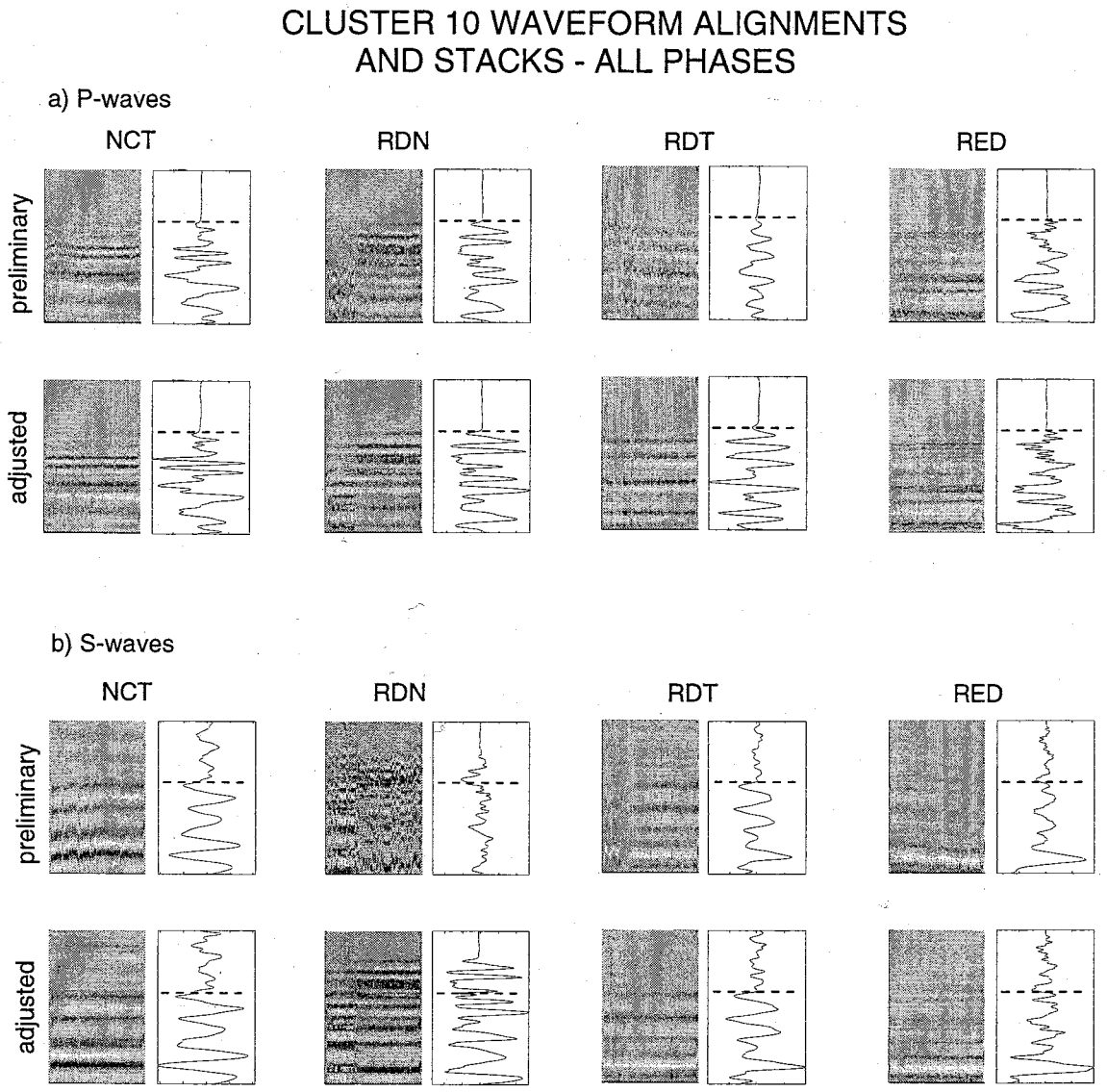


Figure 10: Waveform alignment surfaces for events in the December 13-14 LP swarm. a) P-wave alignments. b) S-wave alignments. Dashed horizontal lines on associated stacks indicate picks on which traces are aligned. The clear trend shown in P-wave preliminary alignments for station RDT and S-wave preliminary alignments for station RED, as well as the large scatter evident for both P and S phases for station RDN, indicate that much of the preliminary location scatter is the result of pick inconsistency. Each panel and stack shows 300 samples (3.0 s) of data.

We illustrate waveform alignment surfaces for this LP cluster in Figures 10a and 10b. Figure 10a displays waveforms aligned on preliminary picks (upper panels) and adjusted picks (lower panels), and corresponding trace stacks for P-waves at the four operating stations; parallel plots are shown for S-wave picks in Figure 10b. A picking bias over time is clearly indicated in P-wave alignments at station RDT (Figure 10a) and S-wave picks at station RED (Figure 10b), where a slope to the preliminary waveform alignments (upper panels) can be seen; note that this trend vanishes when waveforms are aligned on the automatic repicks (lower panels), and strong similarity can be seen among waveforms for the interval displayed.

The waveform alignment slope for P picks at RDT and S picks at RED correlates with a general trend in calculated LP event magnitude through time, suggesting that preliminary phase pick errors for these events are correlated with magnitude. Phase picking bias among smaller events has long been known to be significant and is not only responsible for general completeness thresholds for seismic networks (e.g. Wiemer and Wyss, 2000), but also for thresholds of location confidence and hypocenter bias. In an investigation of shear-wave splitting and seismic anisotropy, Aster et al. (1990, 1991) noted magnitude-correlated biases in phase identification for events in the Anza seismic network, California; however, their discussion suggests that systematic bias may be the product of changes in source-time function, rather than simple noise contamination. Based on the consistent waveforms well into the LP event codas, we believe the bias in the case of these events is primarily a function of inability to detect first arrivals within the background noise for smaller events.

Not only is there greater hypocentral scatter among preliminary event depths (e.g.,

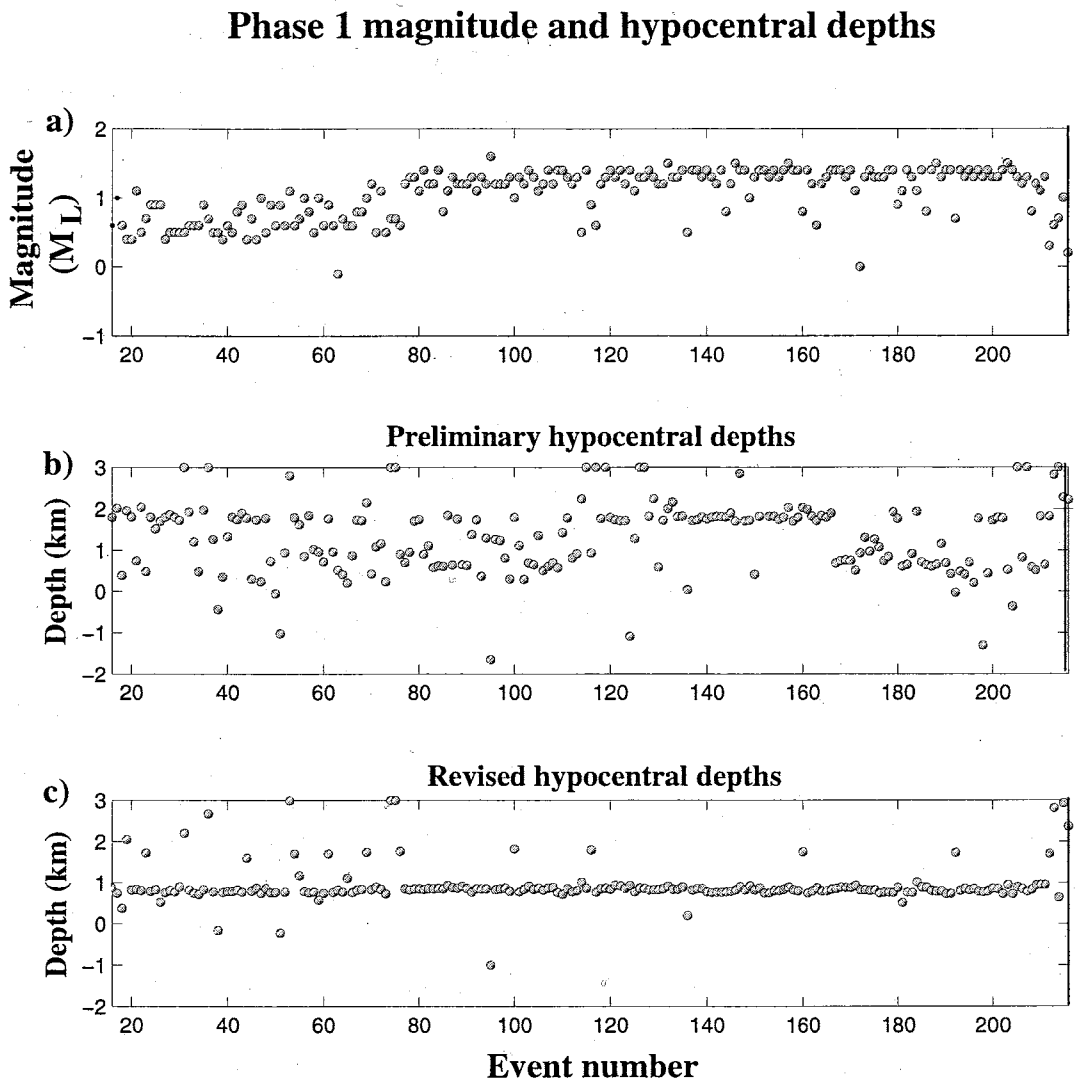


Figure 11: a) Cluster 10 event magnitudes over time. This cluster is dominated by the energetic LP swarm of December 13-14, 1989. b) Preliminary hypocentral depths estimates for these events. c) Hypocentral depth estimates based on relocations after cross-correlation repicking.

Lahr et al., 1994), but also a generally linear correlation exists between magnitude and mean estimated hypocentral depth. In Figure 11a we plot event magnitudes for the time period of 13-14 December, 1989; although there are a few other earthquake types which occurred during this phase, the seismicity is dominated by the shallow LP swarm. Note the steady trend in magnitudes (M_L) for the events, levelling off to a fairly consistent magnitude of 1.2 at event 150 (0620 UTC, 14 Dec.). We plot hypocentral depths as a function of time in Figure 11b and 11c, for preliminary and relocated hypocenters, respectively. The preliminary depths shown in Figure 11b show a clear trend which correlates with event magnitude: the mean depth appears to be shoaling with elapsed time. Relocated hypocentral depths in Figure 11c show that no such mean trend exists among these data, and hypocentral depths are stable throughout the duration of the swarm. The misleading depth trend in Figure 11b suggests that where only preliminary picks are available, interpretation of spatio-temporal seismicity trends in real time may result in erroneous conclusions regarding the evolution of eruption-related activity.

We note that human mispicks for the LP swarm may not be the only contributing factor to the observed scatter in preliminary locations. Figure 12 illustrates waveforms for seven members of the swarm, as recorded at station RED, the only station not equipped with gain-ranging. The dominant, low frequency part of the signals occurs 0.5 to 3.5 s after the high frequency onset of each event. This high-frequency constituent is visible with varying relative strength at all recording stations, but its waveform is less repeatable at the more distant stations from event to event than is the LP portion of the coda. We show waveform alignment plots in Figures 13a and 13b for station RED to illuminate some features we have noted through cross-correlating these events. Figure 13a shows waveform

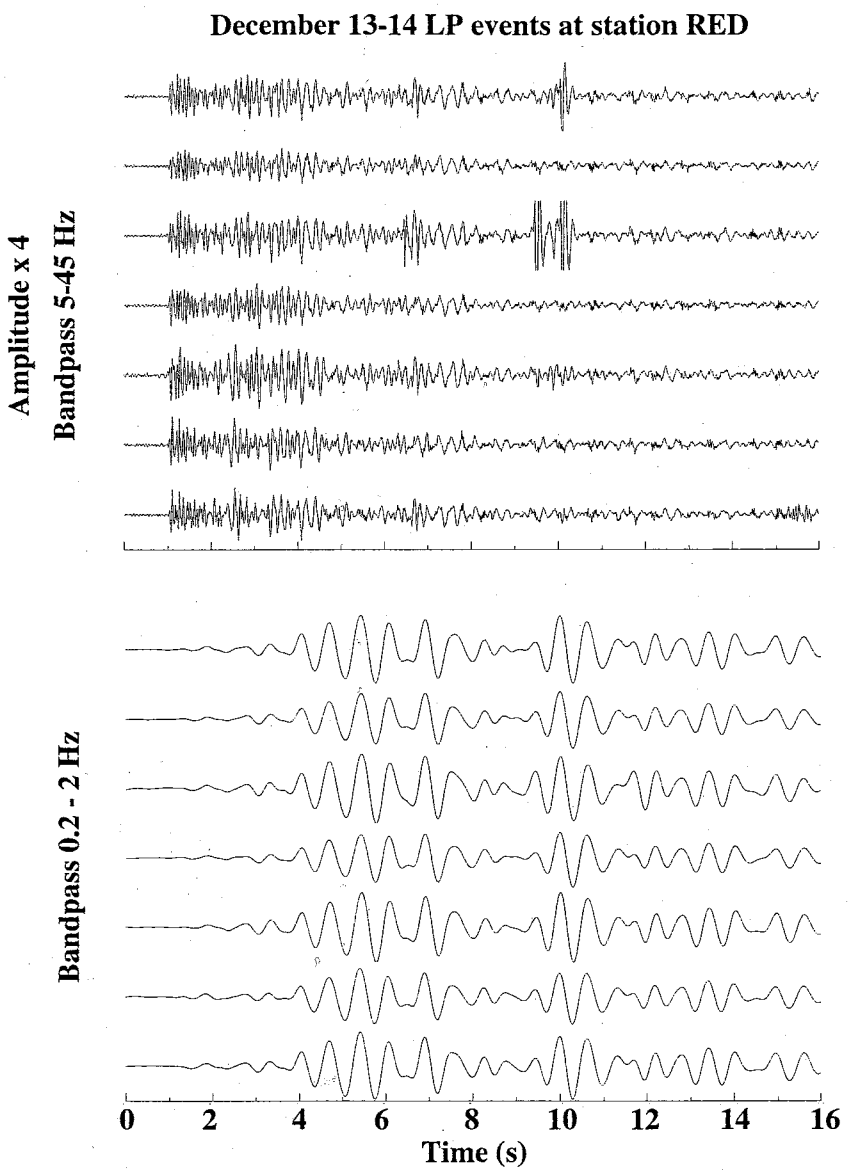


Figure 12: Five representative waveforms for December 13-14, 1989 LP events recorded at station RED. a) Bandpass filtered between 5 Hz and 45 Hz to illuminate the short-period onset of the events. The resulting waveforms suggest that the triggering mechanism for these events may be brittle failure earthquakes, similar to the sources for "hybrid" events, but much smaller with respect to the LP resonance which follows. These amplitudes have been multiplied by a factor of 4 for visibility. b) Bandpass filtered between 0.2 Hz and 2.0 Hz to emphasize the LP portion of the signal, which dominates the traces.

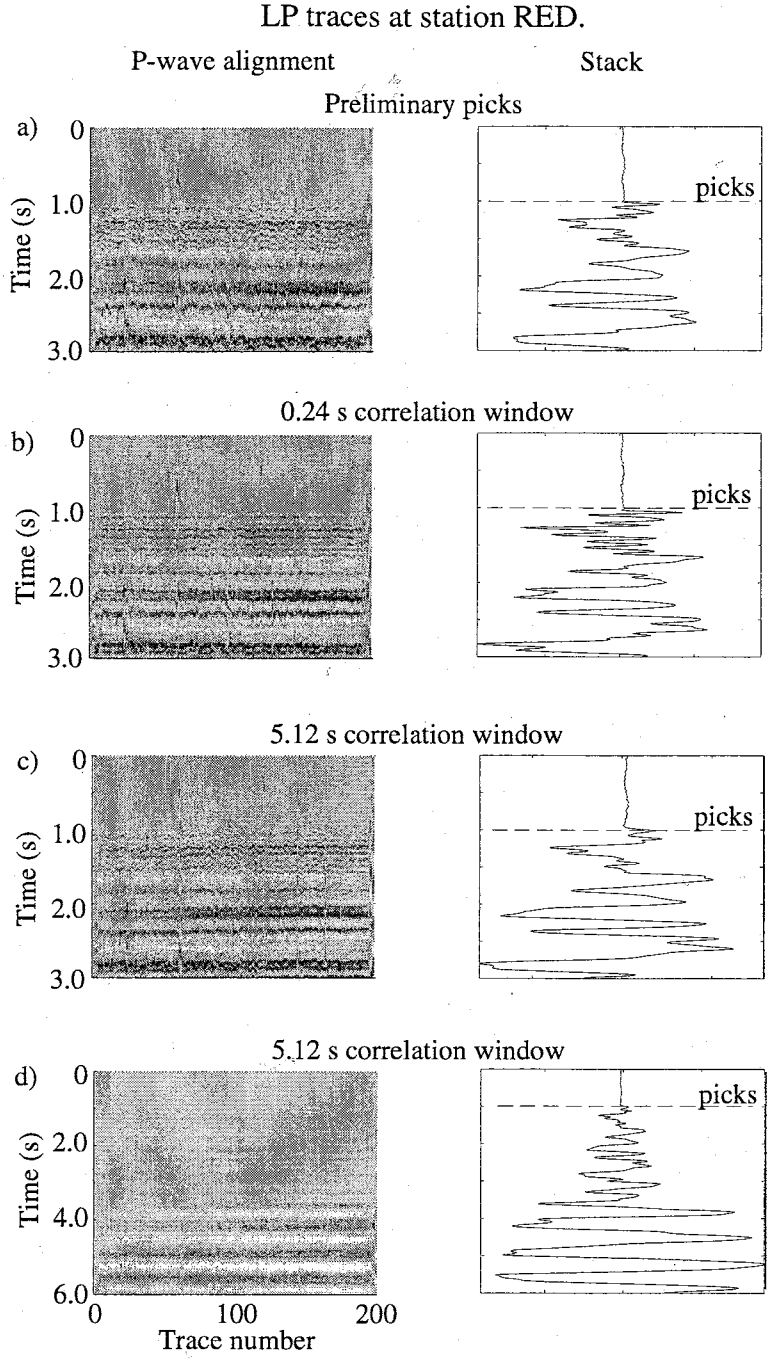


Figure 13: a) RED waveform alignments for the December 13-14 LP swarm. a) alignment on preliminary picks. b) aligned on picks adjusted using a 24-sample (0.24 s) correlation window, to emphasize the high-frequency early component which provides the first arrival.. c) The same traces adjusted using a 512-sample (5.12 s) correlation window to correlate based on the LP portion of the waveform. d) 6 seconds of data for the 5.12 s correlation adjustment, showing that the LP portion of the signal is well-aligned.

alignment and stack for station RED using preliminary picks. In Figure 13b, we have successfully aligned the high-frequency onset by choosing a short (24-sample) correlation window. Significant enhancement is obtained for alignment of higher-frequency constituents of the signal, although the long-period component shows little gain. Examination of the high-pass filtered traces in Figure 12 reveal substantial similarity among the short-period onsets of these events, so the improvement in SP alignment is to be expected when a short windowlength is invoked. Figure 13c shows alignments and stacks for the same waveforms when a longer, 512-sample (5.12 s) correlation window was chosen. Note that no improvement can be seen in the SP portion of the signal - indeed this appears to have degraded considerably in both the waveform alignment and stack plots compared to the original picks of Figure 13a or the 24-sample realignment in Figure 13b. Figure 13d displays six seconds of coda for the long-window realignment, and confirms that the 512-sample correlation has successfully aligned the LP portion of these swarm events, at the expense of the consistency among first arrivals for the SP onset phase. Short-window lag adjustments for these correlations had a maximum range of 0.137 s, whereas the lag adjustments for the longer correlation window spanned 0.2278 s

Preliminary and revised locations of Phase 1 seismicity are illustrated in Figure 14 in map view and east-west cross-section. Preliminary locations for the LP swarm members (red) outline a region of ~2.5 km^3. Figure 14b shows the automatic relocations for Phase 1 events. The LP source region reduces to ~0.02 km^3, consistent with the findings of Lahr et al. (1994). Pick adjustments for each of the correlatable phases for the LP events are shown in Figure 14. Some larger corrections (>0.5 s) are not shown so that we may use a sufficiently fine scale to illustrate most of the data. The greatest range of pick corrections

Phase 1 seismicity

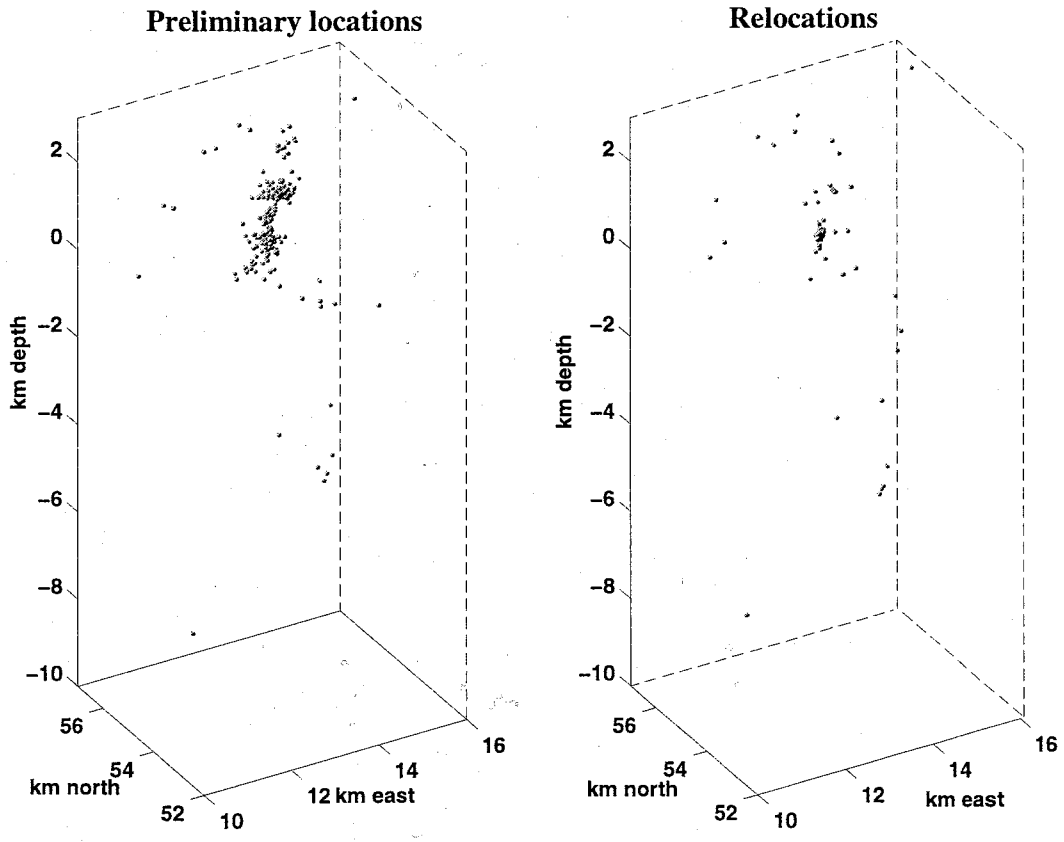


Figure 14: Preliminary and relocated hypocenters for Phase 1 seismicity, which is dominated by the Dec. 13-14, 1989 LP swarm comprising most of Cluster 10. In addition to the collapse of the LP swarm events, a few deep VT events move toward a common location at about 6 km depth; some outlying hypocenters remain unchanged. Mean RMS traveltime error for Phase 1 events reduces from 0.0825 s to.0.0258 upon repicking, with standard deviations of 0.2244 and 0.0398 s for preliminary and repicked hypocenters, respectively.

for P waves is ~0.23 s, and for S waves ~1.21 s. With a shallow velocity layer Vp of about 3 km/s and corresponding Vs of approximately 1.7 km/s, this can account for worst case preliminary mislocations of up to 2 km.

In a discussion of the LP events at Redoubt, Chouet et al. (1994) emphasize the relationship among so-called "hybrid" events, "true" LP events and tremor. Hybrid events are presumed to result from excitation of a resonating fluid-filled crack, where the energy source is a nearby brittle-failure (VT) event or other high-frequency disturbance. LP events, however, are also characterized by a high frequency onset, which is generally only observed at proximal stations because of attenuation. Figure 15 illustrates the three discrete event types, VT, LP and hybrid. It seems probable that the short-period onset of LP events is related to the triggering mechanism for "hybrids," and that the two are genetically indistinguishable. In the case of LP events, the trigger may be near enough to the fluid volume that only a small amount of energy is required to excite resonance, whereas "hybrid" events may represent a more energetic short-period (brittle failure) source, which may therefore occur at a greater distance from the resonator. The high-pass and low-pass filtered trace representations of Figure 12 suggest that the onsets of the December 13-14 LP events are indeed small, brittle-failure microearthquakes. Hence, some preliminary scatter in the LP or hybrid event locations may represent true variation in the locations of the small SP sources which excite the LP resonance. Preliminary analyst picks are determined for event onsets, rather than considering correlation of later portions of event codas. Because these initial SP components of the signal are so small, however, consistent picking of sufficient arrivals to obtain robust hypocenter locations is seldom possible. Locating based on traveltime differences for the later, LP portion of the signal provides the centroid

VOLCANIC EARTHQUAKES RECORDED DURING 1989-1990 REDOUBT ERUPTION. STATION RED - 25 seconds shown.

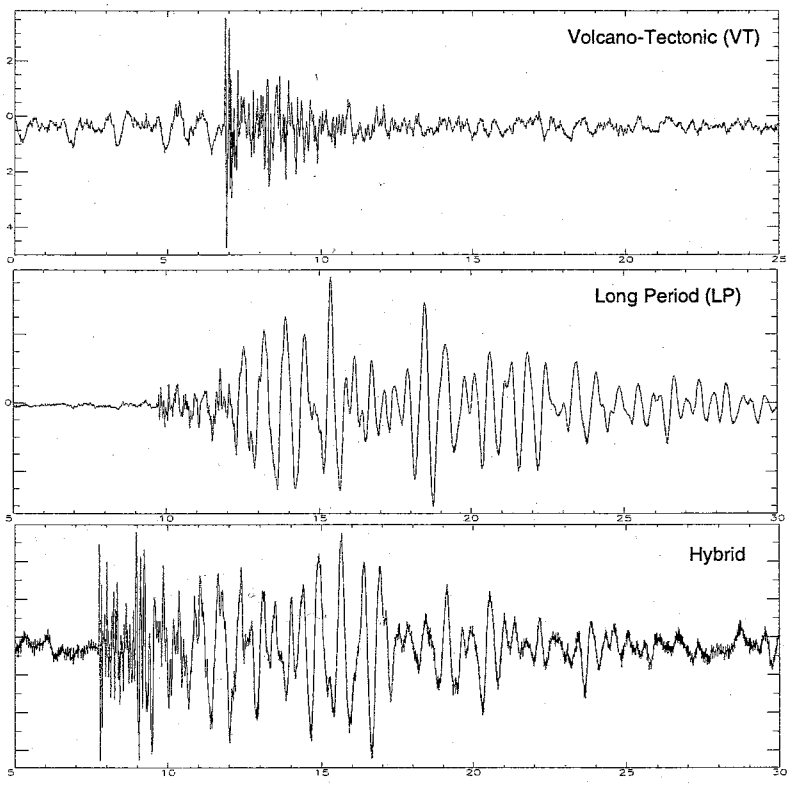


Figure 15: Typical waveforms for volcano-tectonic (VT), volcanic long-period (LP) and "hybrid" type events. LP events (Figure 15b) are dominated by monochromatic, high-amplitude stationary source resonance, but onsets can be seen at nearest stations to contain a high-frequency element. VT events (Figure 15a) result from shear failure of brittle material within or beneath the volcanic edifice, in response to changing local stress regime influenced by magma dynamics. "Hybrid" events (Figure 15c) appear to begin with VT excitations which trigger a LP response in a nearby resonator.

of the volumetric source, but it does not necessarily accurately pinpoint the SP trigger locations.

Phase 1 earthquake locations exhibit mean RMS traveltime errors of 0.0825 s for preliminary phase picks, with a standard deviation of 0.2244 s; relocations using the automatic repicks reduce this mean RMS to 0.0258 s and standard deviation of 0.0398 s. A significant part of the preliminary misfit may be attributed to a few outliers; however, median RMS values for preliminary and relocated Phase 1 events are 0.03 and 0.01 s, respectively. Disregarding gross outliers, then, the overall RMS improvement is still roughly a factor of 120%.

Phase 2: Vent clearing phase (12/14/89-12/19/89)

Following the initial explosion of 14 December, the next five days of activity include individual tephra eruptions on 15, 16 and 19 December. Vigorous volcanic tremor and two swarms of VT earthquakes are reported (Power et al., 1994) during this phase. The AVO waveform catalogue provided 135 events for this period, 128 of which were usable in our analysis. Figure 16 shows preliminary and relocated events for Phase 2, which include some shallow LP and hybrid events as well as two, post-explosion clusters of deeper VT events. Hypocenters, shown in Figure 16, do not show such a dramatic change as for Phase 1 seismicity; however, some previously obscured structures and tightening of clusters are evident. Mean RMS traveltime errors reduce from 0.178 s to 0.058 s, with standard deviations of 1.22 and 0.55 for original and repicked locations, respectively. Median RMS values are 0.05 s and 0.04 s for original and repicked locations; hence, ignoring outliers, the location quality improves by about 20%.

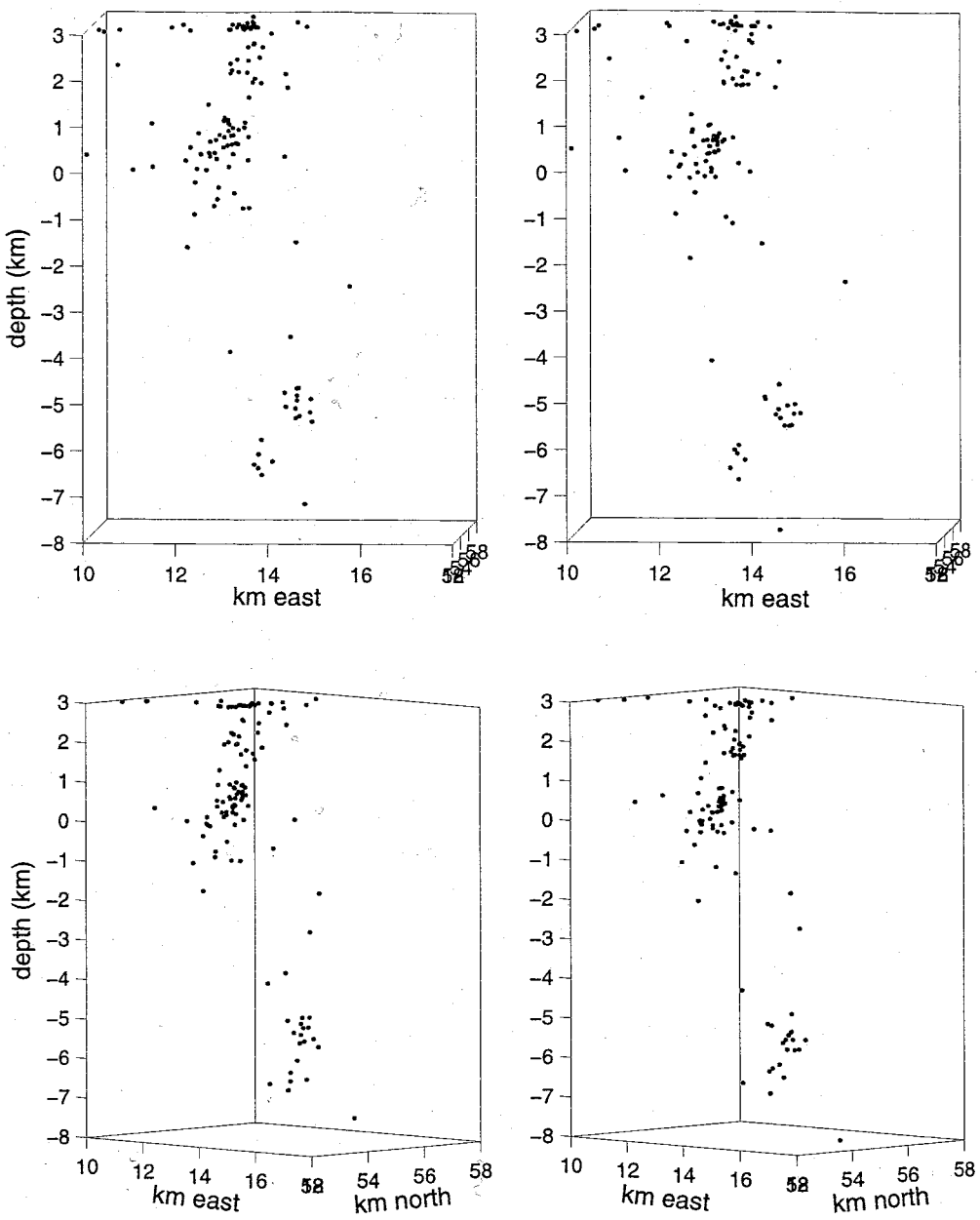


Figure 16: 128 preliminary (left) and relocated (right) hypocenters for Phase 2 seismicity, associated with vent clearing following the initial buildup and first tephra explosion. Some improved clustering can be seen in the hybrid and LP events occurring between 0 and 1 km elevation; two clusters of deeper VT events at 6 km depth exhibit only minor shifts as a result of repicking. RMS traveltime error and standard deviation are 0.178 s and 1.22 s for preliminary locations, and 0.058 s and 0.55 s for relocated events.

Phase 3: First dome-building phase (19 Dec., 1989 - 2 Jan., 1990)

Phase 3 included small, unlocatable shallow events associated with vigorous dome growth, deep (6-10 km) VT earthquakes and shallow LP events which merged into tremor prior to the January 2, dome-destroying tephra explosion (Power et al., 1994). During Phase 3, 395 locatable earthquakes occurred, 353 of which were used in this analysis. Preliminary and relocated hypocenters are shown in Figure 17. Note the generally tighter clusters and sharper delineation of the vertical seismic trend in the relocated events. The two deep VT clusters (at ~5 and ~7 km) collapse in volume, and the suggested lines of hypocenters in the preliminary locations, which occupy depths between ~4.5 and -1 km, sharpen into thin, subvertical lineations upon relocation. The shallow group of LP events at -2 km depth is likewise reduced from a diffuse cloud to a tight cluster. Mean RMS traveltime errors reduce from 0.072 s to 0.043 s, although the median RMS of 0.03 s remains the same for preliminary and repicked locations. RMS standard deviations reduce from 0.238 s to 0.062 s. A plot showing hypocentral depths through time for this phase highlights the behavior prior to the January 2 dome-destroying explosion. Deep VT events dominate the early part of this phase, with two stable clusters at depths of ~6.5 and ~5 km; a gradually shoaling trend of seismicity merges into a re-emergence of shallow activity, which dominates the later part of this phase

Phase 4: Second dome-building phase (2-15 January, 1990)

This phase included three large tephra eruptions on 8, 11 and 16 January, ending in a major dome-destroying eruption on February 15 (Power et al., 1994). Seismicity associated with this activity included VT events and numerous small, shallow events related to dome growth. Much of the small, shallow seismicity was not locatable due in part to the

PHASE 3 SEISMICITY

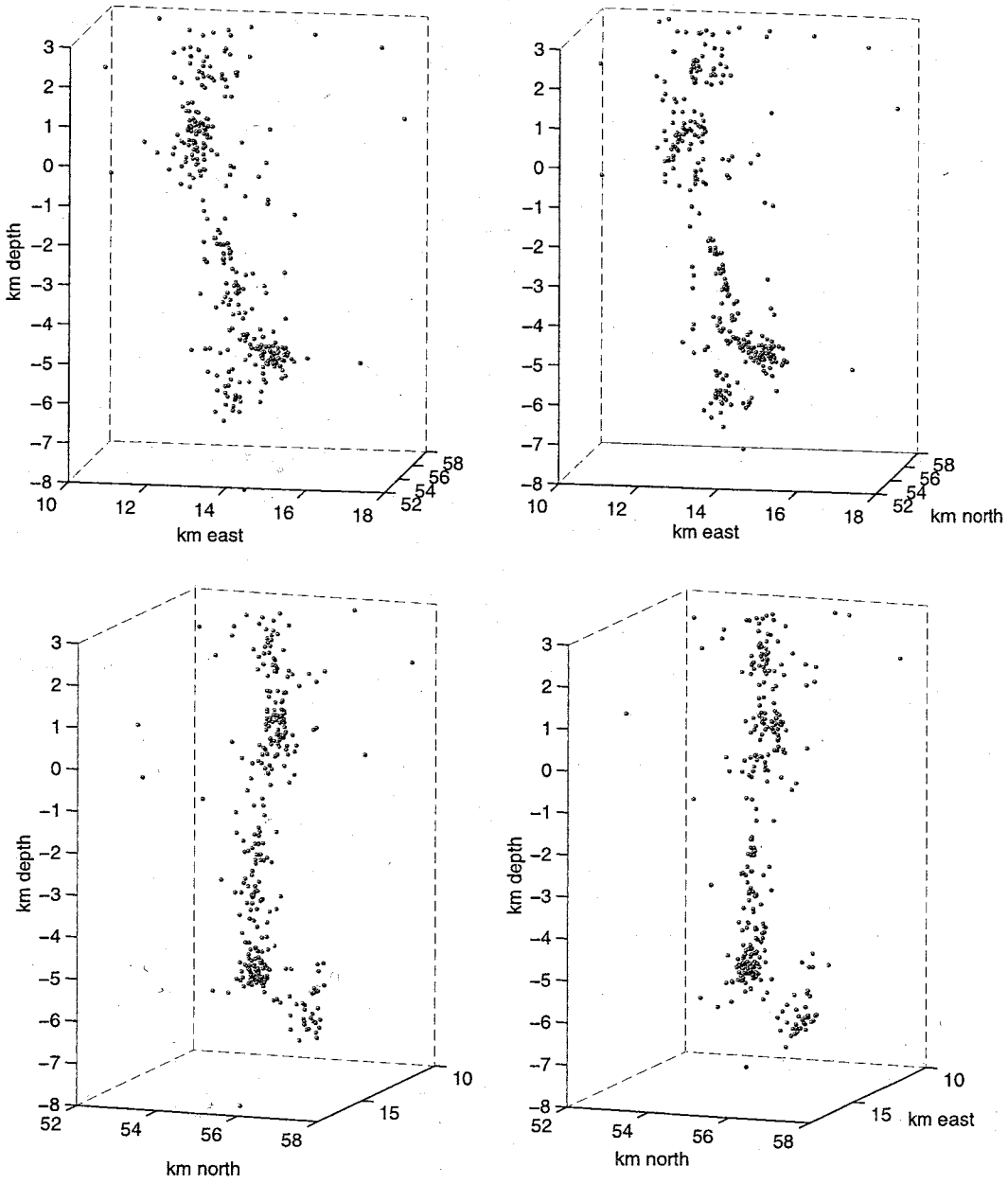


Figure 17: 353 preliminary (left) and relocated (right) hypocenters for Phase 3 seismicity. We show two view angles. looking to the north-northwest (upper views) and to the southwest (lower views). Mean RMS traveltime errors for Phase 3 events is reduced from 0.072 s to 0.043 s, with associated standard deviations of 0.238 and 0.062 s for preliminary and repicked events, respectively.

PHASE 4 SEISMICITY

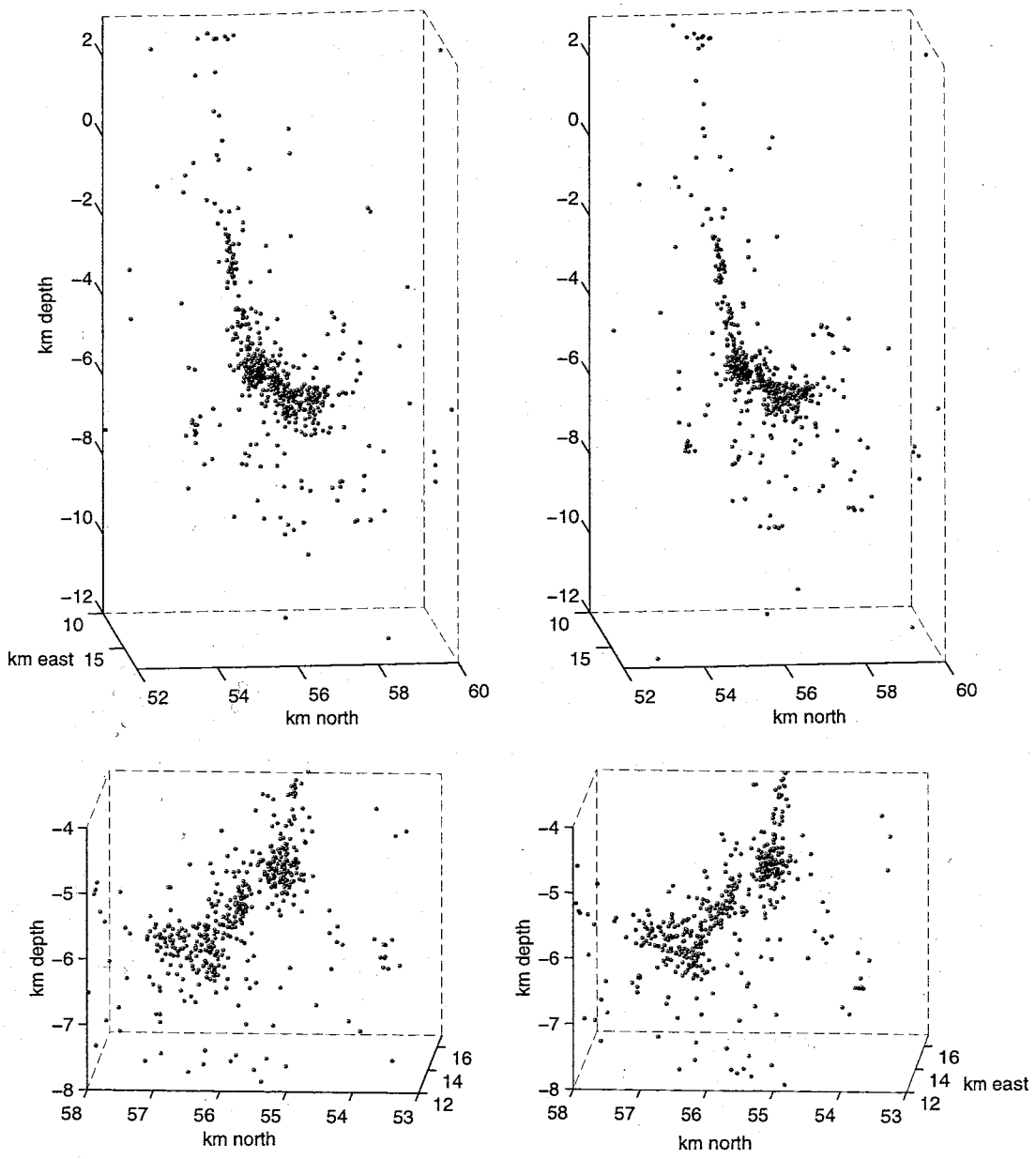


Figure 18: 691 preliminary (left) and relocated (right) hypocenters for Phase 4 seismicity. We show two views of these events; in the upper panels the view is looking west-northwest; in the lower panels we focus on deeper activity (< -4 km) and look towards the east-northeast. Phase 4 event RMS traveltime errors are 0.048 s and 0.0331 s for preliminary and relocated events, respectively, with associated standard deviations of 0.124 and 0.085 s.

January 2 destruction of station RDN by a pyroclastic flow.

723 earthquakes were located during Phase 4; of these we were able to relocate 691. These are comprised almost entirely of VT events. Mean RMS location error reduces from 0.048 s to 0.0331 s, with standard deviations of 0.124 s and 0.085 s for preliminary and repicked locations, respectively. Median location RMS is 0.02 s for both. Figure 18 shows events located during this phase. The subvertical lineation of VT events ranging from the upper portion of the volcano to approximately 5 km depth is more sharply defined after relocation, and its apparent quasi-continuous nature in the preliminary locations is more clearly segmented after repicking. At about 6 km depth, the VT sources become more diffuse, and lineations in relocations of these deeper events appear to outline discrete fractures which may be failing in response to changes in the local stress regime. Figure 18b illustrates a close-up of preliminary and relocated events below 4.8 km depth.

Phase 5: Repetitious Dome-building (15 February - 15 June, 1990)

Overall seismicity levels during this period were significantly lower than in the previous phases (Power et al., 1994). Swarms of small, shallow LP and dome-growth-related activity occurred periodically, but these were by and large too small to locate. Most located events for this four-month phase consist of moderately deep (~ 6km) VT events. We correlated and obtained relocations for 1528 of the 1842 events identified during this period. RMS traveltime errors reduced upon relocation from a mean 0.843 s to 0.593 s, with median values of 0.04 and 0.03 s, respectively. RMS standard deviations went from 0.164 s to 0.127 s.

Figure 19 illustrates preliminary and relocated hypocenters for Phase 5 seismicity. We focus on events deeper than 2 km, as locatable shallow activity during this phase was

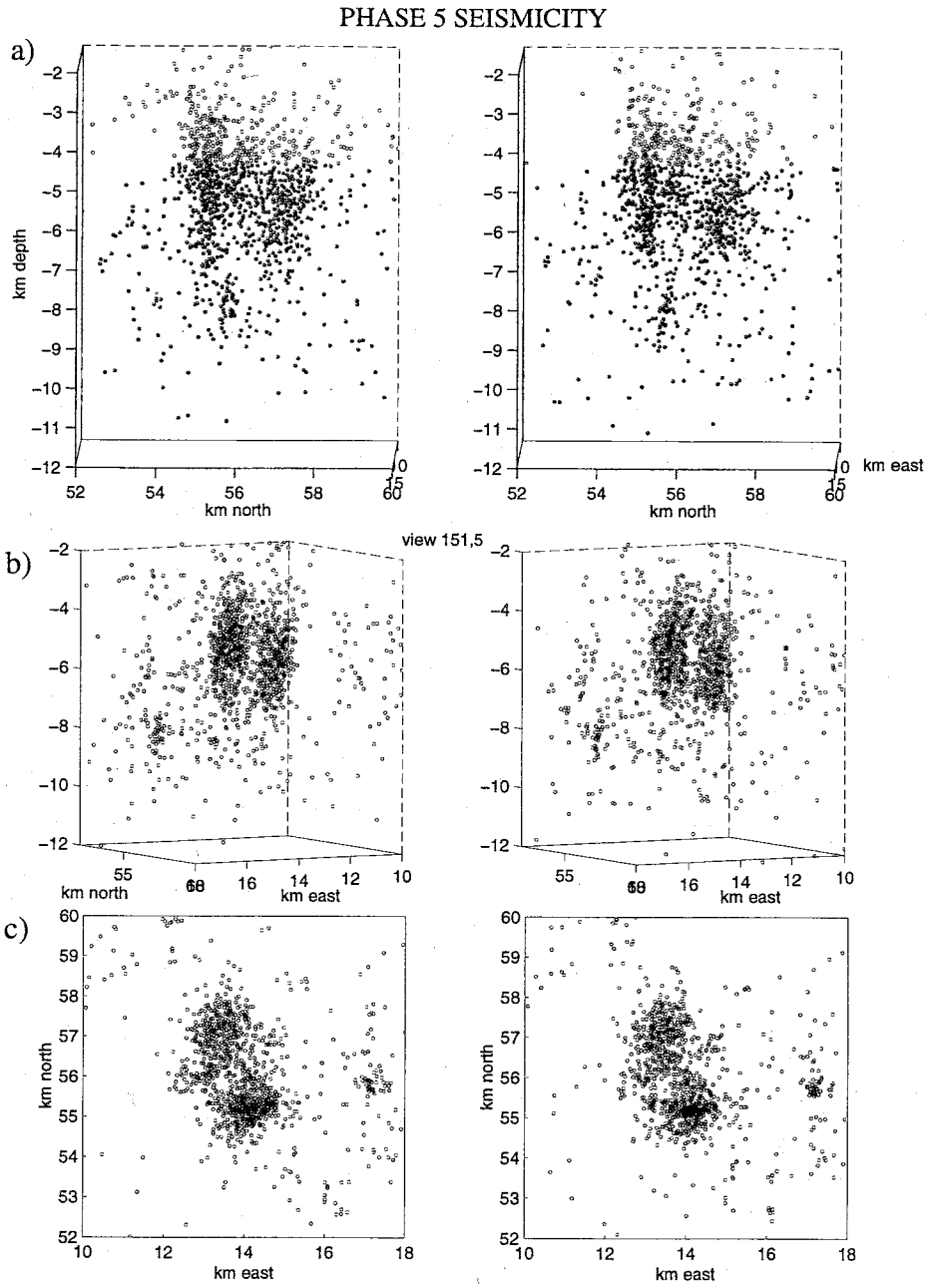


Figure 19: Phase 5 deep seismicity. Preliminary (left) and relocated (right) hypocenters are shown for 1528 events which occurred during this four-month period. a) View looking to the east reveals subvertical lineations in relocated events. b). View looking northeast depicts two distinct volumes of seismicity. c) Map view of deeper seismicity, showing lateral relationship of the two seismogenic regions. Mean RMS values for these events are 0.843 s for preliminary and 0.593 s for revised locations, with standard deviations of 0.164 s and 0.127 s, respectively.

almost nonexistent. There appear to be two general zones of activity centered at about 5km depth, separated by approximately 0.5 km. Figure 19a and 26 b show two cross-sectional views of these zones, showing the suggestion of a linear structure dipping steeply towards the south (Figure 19a) and the horizontal separation of these two seismogenic zones (Figure 19b). Each of these regions occupies about 7.5 km^3. Their centroids are aligned along a strike of about N15°W. Relocations suggest numerous somewhat clustered zones of activity within the two larger seismogenic zones. Figure 19c shows these deeper events in map view.

Phase 6: Post eruption (15 June, 1990 -31 December, 1994)

Locatable post-eruption activity consists almost entirely of VT and tectonic microearthquakes, concentrated near the centroid of deep VT activity of earlier phases, but more diffuse and with considerably diminished intensity. Power et al. (1994) define the end of the post-eruptive phase arbitrarily as October 1, 1990. Figure 20 shows relocations for post-eruptive seismicity; these 1189 earthquakes include activity through December 1994. The bulk of these earthquakes occupy a volume of roughly 36 km^3, centered at a depth of about 5.5 km. Structure within this concentration appears to roughly delineate two sub-parallel tabular bodies which dip to the east at an angle of about 60°.

DISCUSSION

New insights into the eruption processes and seismogenic sources revealed through correlation and relocation invite a closer examination of the eruption-related seismic behavior. In Figures 21-23 we show relocations for the six eruption phases to illustrate the general seismic development through time for the 1989-1990 Redoubt eruption. We

PHASE 6 SEISMICITY

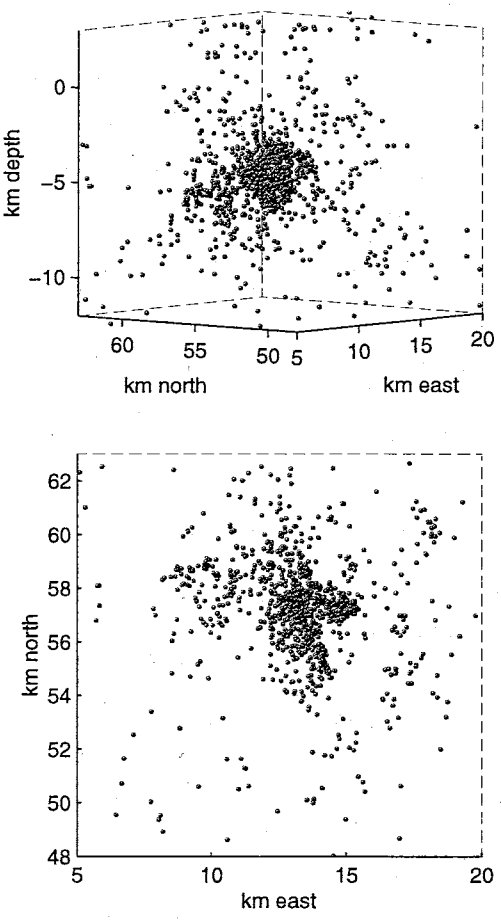


Figure 20: Phase 6 seismicity. Locatable events consisted entirely of VT earthquakes. We show 1189 relocated hypocenters for post-eruptive seismicity, defining a diffuse source region concentrated ~ 5 km below sea level. Some clustering is apparent in the elevation in a), and northwesterly-trending structures are suggested in the map view in b).

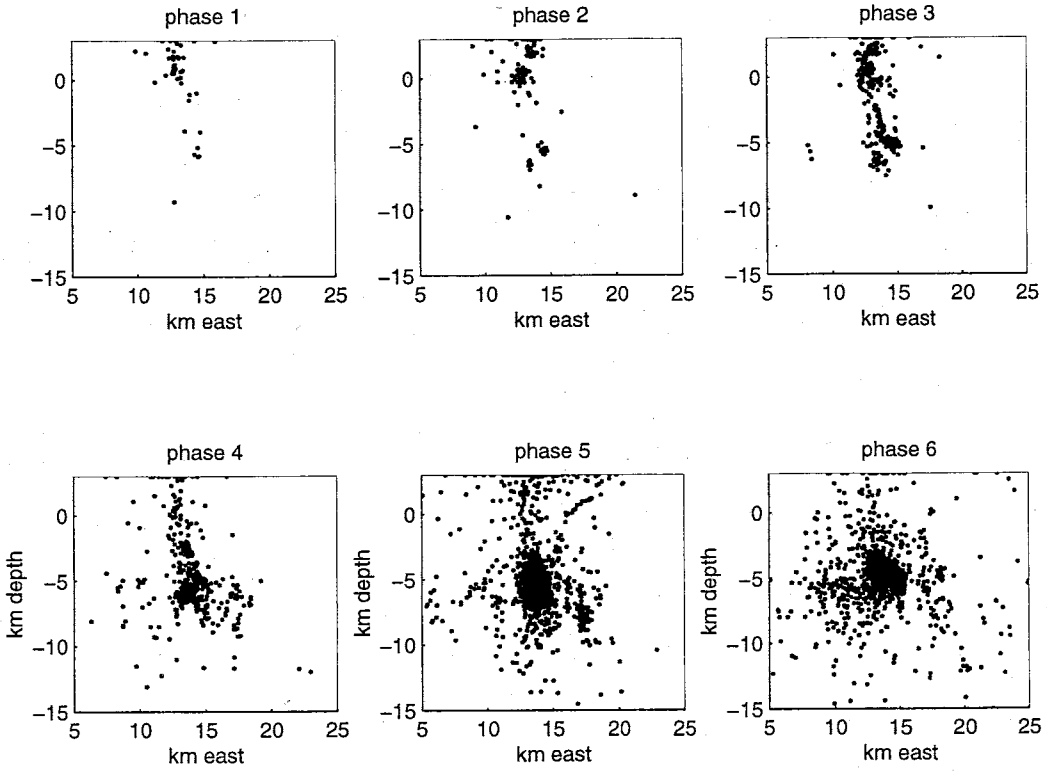


Figure 21: East-west cross-sections of relocated hypocenters for all six eruptive phases. Development and of the seismicity can be followed from the dense cluster of shallow events in Phase 1 through development of the steeply plunging "pipe" of VT events in Phase 3 and increasing deep VT behavior through later eruptive phases.

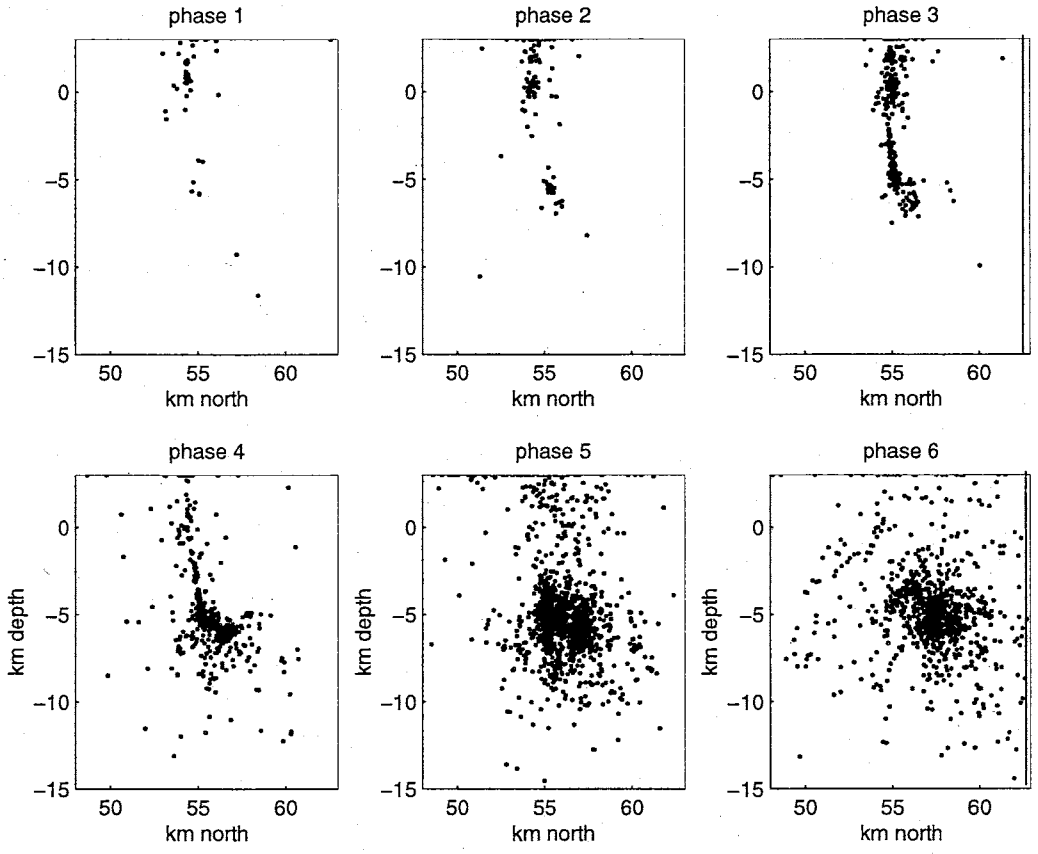


Figure 22: North-south cross-sections of seismicity for the six eruptive phases. Northward shift of the centroid of activity with increasing depth through time is apparent.

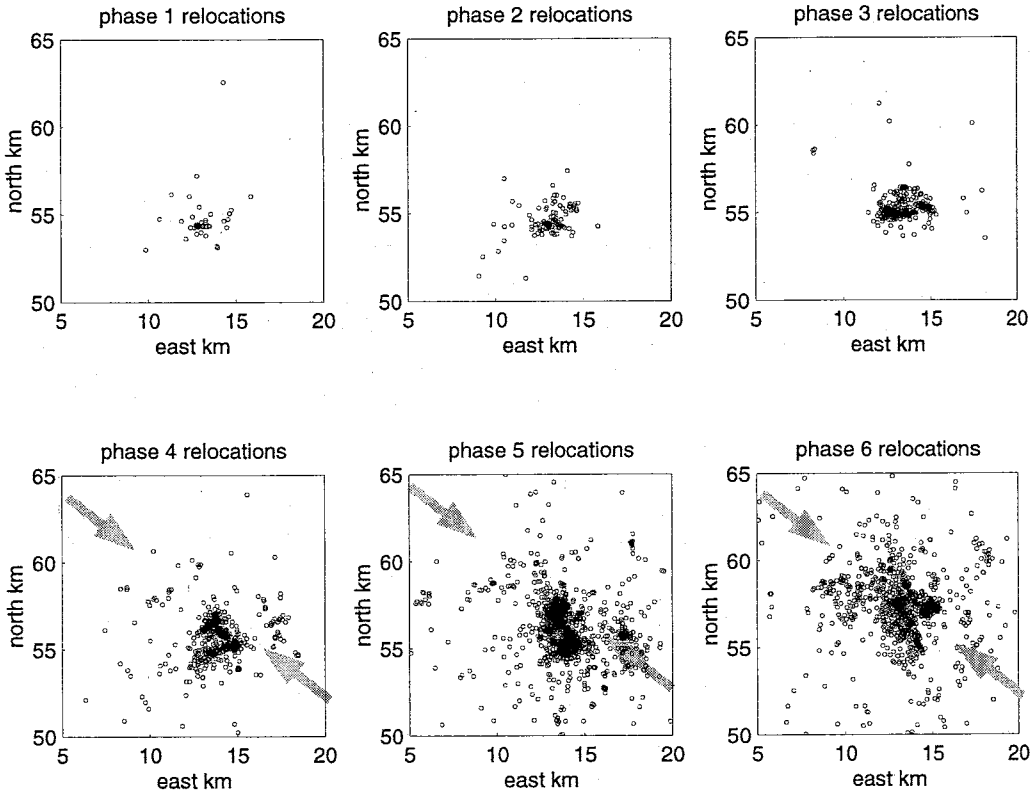


Figure 23: Map view of the seismicity associated with all six eruption phases, illustrates the lateral spread of seismic source area as the eruption progresses. As the subvolcanic structure responds to the changing stress field after evacuation and withdrawal of magma and volatiles, the seismicity moves to more distal locations and becomes less concentrated. This results in a lower proportion of concentrated sources which would have correlatable waveforms. Gray arrows in lower panels indicate approximate direction of maximum horizontal compression due to convergence of the Pacific and North American plates.

show cross-sectional elevations looking north (Figure 21) and eastward (Figure 22), as well as a map view (Figure 23). The temporal development of seismicity clearly indicates outward expansion of the seismogenic volume through, and beyond, the eruptive activity. Some of the systematic trends and temporal changes should be viewed with caution, however, as the perturbation of the local stress field, modifications to the hydrological and local thermal regimes should be expected to modify the seismic velocities in the affected region. Relocating (and, initially, locating) earthquakes using a single, static velocity model may therefore introduce systematic artifacts into the estimated hypocenters which cannot be removed by pick correction alone. Interpretation of some fine details should be approached with caution, however the relocations improve interpretability of some seismogenic features at Redoubt.

During Phase 2, VT events begin following the early eruptions of 14-15 December, and are largely confined to deeper (>5 km) portions of the volcano (Figure 16). This has been interpreted to reflect stress adjustments on deep-seated fractures or faults, following the evacuation of magma and volatiles (e.g. Power et al., 1994). VT activity remains concentrated at depth throughout and beyond the eruption, although a period of shoaling is noted during Phase 3 (Figure 17) between 20 and 26 December. Preliminary locations outline a fairly continuous "pipe" of activity, but relocated events indicate that the seismicity occurs in discrete patches or clumps over the length of the active subvertical feature (Figure 17). This may represent repeated slip on previously existing fractures in the vicinity of the ascending magma column. This behavior may be a response to reduction of normal stress as magmatic volatiles or related circulating hydrothermal fluids elevate the pore pressure in the region immediately below the vent. Upward propagation of this seismicity

may reflect the advance of a fluid pressure front as recharge continues prior to the next explosive phase. This activity was immediately followed by renewed shallow LP swarming.

Persistent deep VT activity during and beyond the eruption (Figures 18-20) implies that there may influences at work other than short-term adjustments to depressurization through removal of magma (e.g., Power et al., 1994). Phase 6 VT events appear to delineate two sub-vertical, intersecting tabular features (Figure 23). These features are oblique to the generally northwesterly direction of subduction in this area. Detailed analysis of shallow crustal stresses for the immediate vicinity of Redoubt is not yet available, but Biswas et al. (1986) estimated principal compressive stress from a shallow earthquake near Mount Spurr (Figure 1) to be oriented N52°W. This direction, indicated as gray arrows in the lower panels of Figure 23, is consistent with the direction of convergence between the Pacific and North American Plates at this location, and approximately the same as motions calculated by Lundgren et al. (1995) using a finite element model to estimate nearby crustal deformation.

Seismicity which has migrated to greater distance from the volcanic center in the latter and post-eruption stages should exhibit behavior that is influenced more by regional tectonic stresses and less by the localized magma-related stress perturbations. Unfortunately these data do not provide sufficient first-motion information to calculate focal mechanisms. The outward diffusion of hypocenters with time results in a reduced event spatial density, which greatly decreases the number of similar earthquakes from which to derive meaningful cross-correlation comparisons; therefore the overall ability to improve the image of small, interpretable structures with our relocation method is reduced for

much of the late and post-eruptive activity. The two intersecting alignments, however, are suggestive of conjugate fracture systems optimally oriented within the regional crustal stress field.

Relocation of the shallow LP events has revealed that the LP source occupies a very small volume (~0.02 km^3), which is consistent with interpretations of Lahr et al. (1994) and Chouet et al. (1994). Closer examination of the earlier, high-frequency portion of the LP waveforms indicates that at least for the 13-14 December swarm, the triggering mechanism itself is a small VT event. We propose that these LP events are by and large therefore genetically indistinguishable from so-called "hybrid" events, and that the primary difference between the two is the relative amplitude of the VT source compared to that of the LP resonance.

Our failure to obtain consistent correlation lags when comparing cross-correlations of the initial, VT portion of these events with the later, LP resonance indicates that the time between arrivals of the VT triggering onset and subsequent LP resonance is inconsistent among swarm events. This represents either some inconsistent time delay in the LP response to the triggering impulse or slight variations in VT distance from (or spatial relationship to) the centroid of the resonating volume.

Repicking and relocation of the LP events has succeeded in removing the temporal depth migration, which was an artifact of preliminary picking inconsistencies. Not only does this verify the stationary nature of the LP source (in this case), but it also highlights the importance of using caution in interpreting apparent temporal seismicity trends during the course of an eruption.

SUMMARY

We have applied a new event classification and repicking algorithm to seismicity associated with the 1989-1990 eruption sequence of Redoubt Volcano, Alaska, to demonstrate earthquake location improvements which may be possible with heterogeneous datasets recorded with small networks in noisy environments. The method is successful in segregating and properly associating volcano-tectonic and volcanic long-period earthquakes, and can provide meaningful fine adjustments to spatial and temporal seismicity trends that may have been misinterpreted due to preliminary location biases. Through application of the technique to this catalogue we have also demonstrated its potential usefulness as a retroactive network diagnostic tool.

The distribution of hypocenters of LP events for the 13-14 December swarm is shown to collapse from a volume of 0.25 km^3 to ~0.02 km^3 after relocation, supporting previous interpretations that the LP resonance emanates from a stationary source. Further, inconsistent time shifts among these events between optimal waveform alignment for the short-period onset versus long-period resonance suggests that there are varying spatial relationships between the triggering sources and the centroid of the resonating volume. A nearly continuous, upward-migrating "pipe" of VT activity is resolved into discrete patches of seismicity, and deeper VT source zones exhibit linear features suggestive of fractures or joints that were previously obscured in the preliminary hypocenter scatter.

ACKNOWLEDGEMENTS

The authors wish to thank the staff of the Alaska Volcano Observatory, particularly John Power (USGS, Anchorage) and Bob Hammond (USGS, Fairbanks), for making

available to us the full triggered waveform catalogue and pickfiles for Redoubt, and Art Jolly (University of Alaska Geophysical Institute, Fairbanks) for help with additional files. This work was supported under a grant from Sandia National Laboratories, Albuquerque, New Mexico and by Niitsuma Laboratories, Tohoku University, Sendai, Japan.

REFERENCES

Aster, R.C., P.M. Shearer and J. Berger, Quantitative measurements of shear wave polarizations at the Anza Seismic Network, southern California: Implications for shear weave splitting and earthquake prediction, *J. Geophys. Res. 95,* 12,449-12,473, 1990.

Aster, R.C., P.M. Shearer and J. Berger, Reply (to Comment on Aster et al., 1990), *J. Geophys. Res. 96*, 6415-6419, 1991.

Aster, R.C. and J. Scott, Comprehensive characterization of waveform similarity in microearthquake data sets, *Bull. Seismol. Soc. Am. 83,* 1307-1314, 1993.

Aster, R.C. and C.A. Rowe, Automatic phase pick refinement and similar event association in large seismic data sets, in *Advances in Seismic Event Location*, edited by C. Thurber and N. Rabinowitz, Kluwer, Amsterdam., 231-263, 2000.

Benz, H.M., B.A. Chouet, P.B. Dawson, J.C. Lahr, R.A. Page and J.A. Hole, Three-dimensional P and S wave velocity structure of Redoubt Volcano, Alaska, *J. Geophys. Res. 101*, 8111-8128, 1996.

Biswas, N.N., K. Aki, H. Pulpan and G. Tytgat, Characteristics of regional stresses in Alaska and neighboring areas, *Geophys. Res. Lett. 13*, 177-180, 1986.

Chouet, B., New methods and future trends in seismological volcano monitoring, in *Mon-*

itoring and Mitigation of Volcano Hazards, edited by R. Scarpa and R.I. Tilling, Springer-Verlag, Berlin: 23-97, 1996a.

Chouet, B., Long-period volcano seismicity: its source and use in eruption forecasting, *Nature 380*, 209-316, 1996b.

Chouet, B.A., R.A. Page, C.D. Stephens, J.C. Lahr and J.A. Power, Precursory swarms of long-period events at Redoubt Volcano (1989-1990), Alaska: Their origin and use as a forecasting tool, *J. Volc. Geotherm. Res. 62*, 95-135, 1994.

Dawson, P.B., B.A. Chouet, J.C. Lahr, R.A. Page, J.R. VanSchaack and E.E. Criley, *Data Report for a Seismic Study of the P and S Wave Velocity Structure of Redoubt Volcano, Alaska*, USGS Open-File Report 96-703, 43 pp, 1996.

Dodge, D.A., G.C. Beroza and W.L. Ellsworth, Foreshock sequence of the 1992 Landers, California earthquake and its implications for earthquake nucleation, *J. Geophys. Res. 100,* 9865-9880, 1995.

Fehler, M., W.S. Phillips, L. House, R.H. Jones, R. Aster and C. Rowe, Improved relative locations of clustered earthquakes using constrained multiple event location, *Bull. Seismol. Soc. Amer 90*, 775-780, 2000.

Fremont, M.-J. and S. D. Malone, High precision relative locations of earthquakes at Mount St. Helens, Washington, *J. Geophys. Res. 92*, 10,223-10,236, 1987.

Gillard, D., A.M. Rubin and P. Okubo, Highly concentrated seismicity caused by deformation of Kilauea's deep magma system, *Nature 384*, 343-346, 1996.

Goldstein, P., D. Dodge, M. Firpo, and S. Ruppert, What's new in SAC2000? Enhanced processing and database access, *Seismol. Res. Lett. 69*, 202-205, 1998.

Got, J.-L., J. Frechet and F.W. Klein, Deep fault plane geometry inferred from multiplet

relative relocation beneath the south flank of Kilauea, *J. Geophys. Res. 99*, 15,375-15,386, 1994.

Harris, M. and C. Young, MatSeis: a seismic GUI and tool-box for MATLAB, *Seismol. Res. Lett., 68*, 267-269, 1997.

Ito, A., High resolution relative hypocenters of similar earthquakes by cross-spectral analysis method, *J. Phys. Earth 33*, 279-294, 1985.

Jones, R.H. and R.C. Stewart, A method for determining significant structures in a cloud of earthquakes, *J. Geophys. Res. 102*, 8245-8254, 1997.

Kissling, E., Geotomography with local earthquake data, *Rev. Geophys. 26*, 659-698, 1988.

Lahr, J. C., *HYPOELLIPSE / Version2.0: A computer program for determining local earthquake hypocentral parameters, magnitude, and first motion pattern*, USGS. Open File Report 89-116, 92 pp., 1989.

Lahr, J.C., B.A. Chouet, C.D. Stephens, J.A. Power and R.A. Page, Earthquake classification, location and error analysis in a volcanic environment: implications for the magmatic system of the 1989-1990 eruption of Redoubt Volcano, Alaska, *J. Volc. Geotherm. Res. 62*, 137-152, 1994.

Lance, G.N. and W.T. Williams, A general theory for classificatory sorting strategies 1. Hierarchical systems., *Computer Journal 10*, 271-276, 1967.

Lee, W.H.K., The Xdetect program, a course on P_-based seismic networks. In: W.H.K. Lee and D.A. Dodge (Editors), *U.S. Geological Survey Open-File Report 92-441,* 138-151, 1992.

Lees, J. M. and R.S. Crosson, Tomographic inversion for three-dimensional velocity struc-

ture at Mount St. Helens using earthquake data, *J. Geophys. Res. 94,* 5716-5728, 1996.

Lorenzen, R.J.L., *Seismic Tomography of Redoubt Volcano*, M.S Thesis, University of Alaska Fairbanks, 78 pp, 1994.

Ludwig, J.A. and J.F. Reynolds, *Statistical Ecology, a Primer on Methods and Computing,* John Wiley & Sons, New York, NY, 337pp., 1988.

Luetgert, J.H., User's manual for RAY84/R83PLT, interactive two-dimensional raytracing/synthetic seismogram package, *U.S.G.S. Open-file Report 88-0238,* 52 pp, 1988.

Lundgren, P., Saucier, F., Palmer, R. and Langon, M., Alaska crustal deformation: Finite element modeling constrained by geologic and very long baseline interferometry data. *J. Geophys. Res. 100*, 22033-22046, 1995.

Mori, J.J., D. Eberhart-Phillips and D. H. Harlow, Three-dimensional velocity structure at Mount Pinatubo; resolving magma bodies and earthquake hypocenters, in: *Fire and mud; eruptions and lahars of Mount Pinatubo, Philippines*, Philippine Institute of Volcanology and Seismology, Quezon City, Philippines | University of Washington Press: 371-382, 1996.

Pearson, C.F., Seismic Refraction Study of Augustine Volcano, M.S. Thesis, University of Alaska Fairbanks, 131 pp, 1977.

Phillips, W.S., Precise microearthquake locations and fluid flow in the geothermal reservoir at Soultz-sous-Forets, France, *Bull. Seismol. Soc. Am. 90*, 212-228, 2000.

Poupinet, G., W.L. Ellsworth and J. Frechet, Monitoring velocity variations in the crust using earthquake doublets: An application to the Calaveras Fault, California, *J.*

Geophys. Res. 89, 5719-5731, 1984.

Power, J.A., March G.D., Lahr, J.C., Jolly, A.D., and Cruse, G.R., Catalog of earthquake hypocenters at Redoubt Volcano and Mt. Spurr, Alaska: October 12, 1989 - December 31, 1990: *U.S. Geol. Surv. Open-File Report, 93-685-A*, 57 pp., 1993.

Power, J.A., J.C. Lahr, R.A. Page, B.A. Chouet, C.D. Stephens, D.H. Harlow, T.L. Murray and J.N. Davies, Seismic evolution of the 1989-1990 eruption sequence of Redoubt Volcano, Alaska, *J. Volc. Geotherm. Res. 62*, 69-94, 1994.

Pujol, J., Joint hypocentral location in media with lateral velocity variations and interpretation of the station corrections, *Phys. Earth Planet. Inter. 75*, 7-24, 1992.

Pujol, J. and R.C. Aster, Joint hypocentral determination and the detection of low-velocity anomalies; an example from the Phlegraean fields earthquakes, *Bull. Seismol. Soc. Amer. 90*, 129-139, 1990.

Rogers, J.A., S. Maslak and J.C. Lahr, A seismic electronic system with automatic calibration and crystal reference, *U.S. Geol. Surv. Open-File Report 80-324*, 130 pp, 1980.

Rowe, C.A. and J.N. Davies, Analysis of continuous digital seismic records for the 1989-1990 Redoubt Volcano eruption, *EOS Trans. Am. Geophys. U. 1709*, 1990.

Rowe, C.A., R.C. Aster, B. Borchers and C.J. Young, Relocation of seismicity associated with the 1989-1990 eruption of Redoubt Volcano, Alaska. Part i: An automatic, adaptive algorithm for refining phase picks in large seismic data sets, *J. Geophys. Res*, in prep, 2000a.

Rowe, C.A., R.C. Aster, W.S. Phillips, R.H. Jones and M.C. Fehler, Relocation of induced microseismicity at the Soultz geothermal reservoir using automated, high-preci-

sion repicking, *Pure Appl. Geophys.*, submitted, 2000b.

Rubin, A. M., D. Gillard and J.-L. Got, A reinterpretation of seismicity associated with the January 1983 dike intrusion at Kilauea Volcano, Hawaii, *J. Geophys. Res. 103*, 10,003-10,015, 1998.

Rubin, A.M., D. Gillard and J.-L. Got, Microseismic lineations along creeping faults, Nature, submitted, 1999.

Rubin, A.M., D. Gillard and J.-L. Got, Streaks of microearthquakes along creeping faults, *Nature 400*, 635-641, 1999.

Scarpa, R., Seismic tomography and modelling of complex geological structures, *J. Appl. Geophys. 30*, 119-130.

Sneath, P.H.A. and R.R. Sokal, *Numerical Taxonomy*, Freeman, San Francisco, CA., 1973.

Stephens, C.D., B.A. Chouet, R.A. Page, J.C. Lahr and J.A. Power, Seismological aspects of the 1989-1990 eruptions at Redoubt Volcano, Alaska: the SSAM perspective, *J. Volcanol. Geotherm. Res. 62*, 153-182, 1994.

Sonafrank, G.H.C., J. Power, G. March and J. Davies, Acquisition and automatic processing of seismic data at the Geophysical Institute, University of Alaska, *Seismol. Res. Lett. 62*, 23, 1991.

Wiemer, S. and M. Wyss, Minimum magnitude of completeness in earthquake catalogs: Examples from Alaska, the western United States and Japan, *Bull. Seismol. Soc. Amer. 90*, 859-869, 2000.

Part C: Relocation of Induced Microearthquakes at the Soultz Geothermal Reservoir Using Automated, High-Precision Re-picking

ABSTRACT

An automatic, adaptive, correlation-based algorithm for adjusting phase picks in large digital seismic data sets provides significant improvement in seismicity resolution using only a small fraction of the time and manpower which would be required to re-analyze waveforms manually or semi-automatically. The technique includes innovations to standard signal processing approaches, as well as clustering and cluster stacking methods to address both intra-cluster and inter-cluster relationships. We apply this technique first to a small, previously manually repicked subset of the Soultz catalogue so that we may compare our results to those obtained from painstaking visual cross-correlation-based techniques; we then apply the method to a ~7,000-event dataset of Soultz microseismicity from September 1993. The results using the automatic phase repicking show substantial improvement in both P-wave and S-wave arrival estimation consistency. Percentage of events clustering within 5 m of their nearest neighbor increases from 5% to 26% of the catalogue. Hypocenter relocations delineate narrow, linear features previously obscured within the seismic cloud, interpreted as faults or fractures which correspond to fluid propagation paths. RMS traveltime residuals for the larger dataset were reduced by only 0.2%; however, phase pick biases in the preliminary catalogue have influenced both the velocity model and station correction calculations, which will affect location residuals. These pick

biases are apparent on the adjusted, stacked waveforms and may be corrected prior to refining velocity and station correction parameters.

INTRODUCTION

The Soultz-sous-Forêts geothermal field is located in the Alsace region of eastern France, on the western edge of the Rhine Graben (Figure 1), about 50 km north of Strasbourg (Baria et al., 1999). It is situated within the former Pechelbron oil field, in an area of high subsurface temperature gradients (Baria et al., 1999). The target geothermal reservoir resides within granitic rock between 2 to 4 km depth (Baria et al., 1999). This granite has been studied in detail using high-resolution borehole imaging and has been found to host two distinct populations of hydrothermally sealed fractures, striking N20°E and N30°W, respectively (Genter and Traineau, 1996). Stress estimates at Soultz indicate that the maximum horizontal stress within the reservoir is oriented roughly N10°W (Jung, 1991; Cornet and Jones, 1994); hence many observed fractures strike sub-parallel to the maximum horizontal stress and are favorably oriented for hydraulic stimulation of the reservoir (Baria et al., 1999). Hydrofracturing at the reservoir in September, 1993 resulted in over 16,000 microearthquakes, recorded at three deep, down-hole multicomponent seismometers and one down-hole hydrophone (Figure 2). Numerous studies have been undertaken using Soultz microearthquakes to estimate growth of the newly fractured portion of the rockmass (Jones et al., 1995), fracture orientations and flowpaths (e.g. Phillips, 2000; Starzec et al., 2000).

Interpretation of seismogenic features depends critically upon the accuracy and interpretability of the source regions illuminated by the estimated hypocenters. In typical

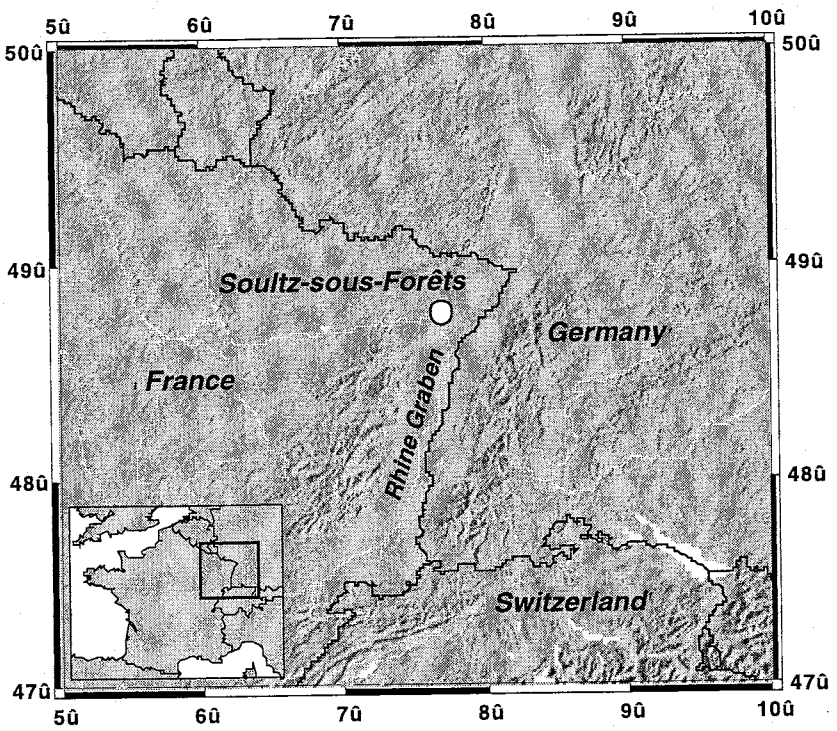


Figure 1: Location map for the Soultz-sous-Forets reservoir (from Phillips, 2000), situated on the western edge of the Rhine Graben in eastern France. This valley is an extensional rift exhibiting high geothermal heat flow.

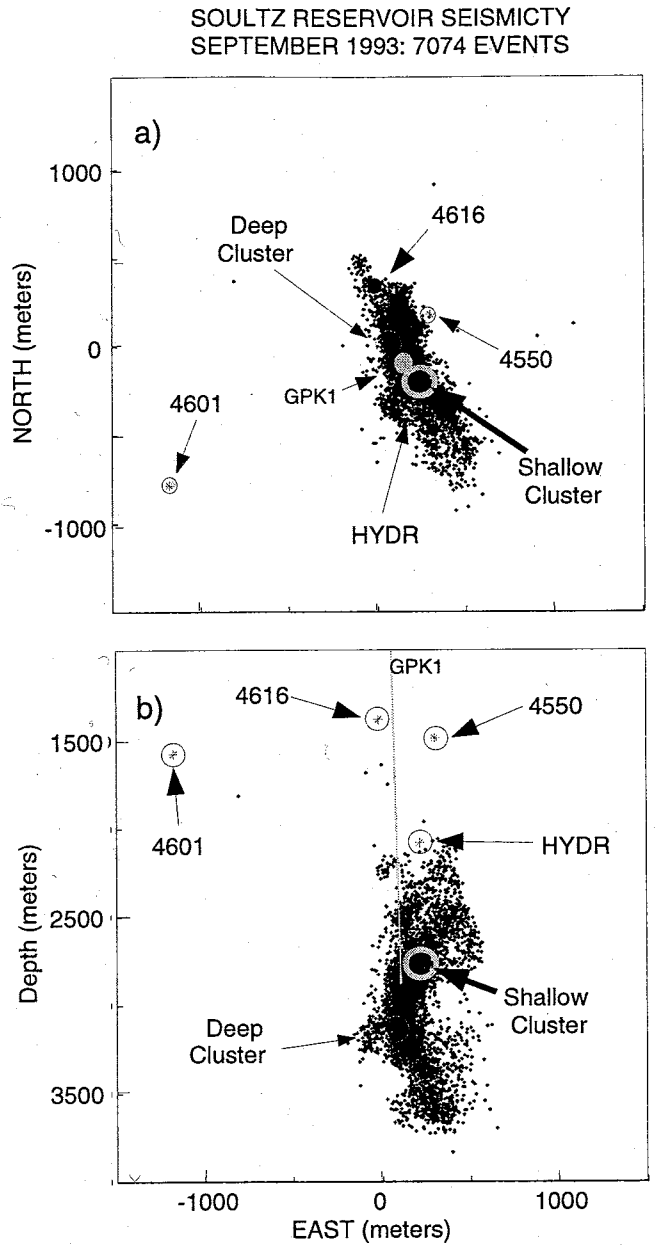


Figure 2: Soultz sous Forets geothermal reservoir (after Aster and Rowe, 2000). Shown are the four borehole sensors (multicomponent seismometers 4550, 4616 and 4601, and hydrophone HYDR) in both map view (a) and east-west cross-section (b). The cloud of preliminary hypocenter locations is shown, along with the position of injection well GPK1. Gray circles indicate approximate locations for the shallow and deep clusters (Phillips, 2000), which were hand-repicked in earlier work.

microearthquake studies at a local or regional scale, hypocenter mislocations of up to several kilometers may be tolerated, when only the identification of general seismic source regions and trends (and associated possible macroseismic effects) are the primary concern. Significant improvement in event locations for local and regional networks has been demonstrated through improved velocity modelling via joint hypocenter determination (JHD) (e.g. Pujol, 1992) or local three-dimensional tomography (e.g. Kissling, 1988; Block, 1991; Block et al., 1994).

Earthquake location accuracy, however, depends not only upon the characterization of the seismic velocity structure through which seismic body waves have travelled, but also upon the ability to estimate accurately the arrival times of those waves. The most common human or computer picking approach is done one event at a time, for many recording stations. Slight to moderate picking inconsistencies are manifested ubiquitously at any particular station among even very similar events, resulting from variations in signal to noise ratio and other effects, as well as human subjectivity. Although one source of systematic mislocation can be successfully addressed through improved velocity modelling, fine details of the seismogenic structures are still commonly blurred by the intrinsic "fuzziness" of hypocenters arising from the remaining random pick errors and associated travel-time residuals (Jones and Stewart, 1997). The picking inconsistencies and resulting hypocentral scatter likewise affect other applications which rely upon the parametric data and associated error estimates. Such applications as reciprocal array processing (Spudich and Bostwick, 1987), spatial mapping of frequency-magnitude distribution (e.g., Wiemer and Wyss, 1997), and evolution of fractal dimension (e.g. Nanjo et al., 1998), as well as general understanding of spatio-temporal seismicity trends in many contexts, will benefit

from correction of large numbers of picks. Certainly the reduction of pick scatter is crucial to mapping of very small-scale structures such as the fracture networks within a geothermal field (e.g., Roff et al., 1996; Phillips et al., 1997), repeating earthquakes (e.g. Nadeau et al., 1995) or deep fault drilling (e.g., Thurber et al., 2000), where precision to a scale of a few meters or less is desired.

A statistical approach to addressing hypocenter fuzziness caused by pick inconsistencies has been presented by Jones and Stewart (1997) and has been incorporated into JHD inversion by Fehler et al. (2000). This technique, referred to as "collapsing," is based on assumptions of Gaussian pick errors and the observation that randomly perturbed points spread out in space. Hypocenters are permitted to move towards an event-based center of gravity, constrained by the dimensions of their error ellipsoids. Although collapsing can be prone to artifacts, synthetic tests as well as comparison to results from exhaustive, manual repicking of real data sets have demonstrated considerable success (e.g., Jones and Stewart, 1997; Fehler et al., 2000). This suggests that, in the absence of waveform data which might permit correction of the picks, collapsing, in combination with techniques for adjusting the velocity model and/or near-surface static corrections, may be a good approach to improve delineation of seismogenic features.

It is possible, however, to address picking inconsistencies directly when digital waveform data are available, using signal processing methods to adjust the picks for consistency among events exhibiting waveform similarity (e.g., Dodge et al., 1995; Shearer 1997, 1998; Rubin et al., 1998, 1999; Aster and Rowe, 2000). Subsequent hypocenter relocation may then proceed using either single-event location, JHD, or simultaneous velocity and hypocenter inversion. Sharpening of the seismogenic source volume image has

been demonstrated through relative relocation using waveform cross-correlation techniques (e.g. Fremont and Malone, 1987; Deichmann and Garcia-Fernandez, 1992; Got et al., 1994; Nadeau et al., 1995; Gillard et al., 1996; Shearer, 1997, 1998; Lees, 1998; Rubin et al., 1998; Rubin et al., 1999;) or combinations of differential cross-correlation and catalogue traveltimes (Waldhauser et al., 1999). In these studies, typical errors of 1 km or more have been reduced to relative location errors of a few tens of meters or less, allowing previously undetected structures within the seismic source zone to be observed. Our method (Aster and Rowe, 2000; Rowe et al, 2000) addresses the correction of scatter which arises from inconsistent picking, by using estimated relative pick lags to calculate corrected picks for consistency among similar waveforms (e.g., Dodge et al., 1995) without imposing constraints from preliminary catalogue parameters on our corrected picks.

DATA

The stimulated volume of the Soultz reservoir occupies roughly 0.48 km^3 in a depth range of about 2100 to 3800 m. Median distances from microearthquakes to the nearest and most distant receiver were approximately 800 m and 2100 m, respectively. Data were digitized at a sampling rate of 5 KHz. The borehole network consisted of three, 4-component (redundant) sensors and one hydrophone. Four-component data were projected into three orthogonal components prior to our analysis.

Location of the Soultz microearthquakes from preliminary analyst picking resulted in the seismicity cloud shown in Figure 2, consisting of hypocenters of ~7,000 events which will be used in our analysis. This subset of microearthquakes was chosen from the larger catalogue based on the availability of seven phase picks from preliminary analysis. Unlike relocation efforts in large datasets recorded by large regional seismic networks

(e.g., Shearer, 1997, 1998; Waldhauser et al., 1999), the Soultz data present significant additional location difficulties resulting from the sparse, four-station network. In the case of Soultz, all available phases are needed to adequately constrain hypocenters, and it was therefore not advisable to neglect arrivals which were noisy or nodal (Phillips, 2000). Further, the sparse network geometry (Figure 2) can cause asymmetric errors due to the location of the most distant station (4601), which provides much of the east-west location control.

Preliminary hypocenters determined using single event location (SEL) methods with initial arrival time picks and velocities determined by Jupe et al. (1994) are shown in Figure 2. These events were located with four P-wave and three S-wave arrivals (no S was available from the hydrophone channel, as S is generally unreadable there). A uniform P-wave velocity of 5.85 km/s and S-wave velocity of 3.34 km/s were used, consistent with data from a shot fired in well GPK1 at a depth of 3360 m. Individual station corrections determined from the calibration shot and applied for SEL hypocenters (Jupe et al., 1994), and station corrections determined by a recent JHD inversion, are shown in Table 1.

Table 1: Station corrections (ms) for SEL and JHD Soultz microearthquake locations.

Station	SEL P	SEL S	JHD P	JHD S
4550	0.0	0.0	-5.49	-8.25
4601	-17.0	-19.0	-16.4	-18.9
4616	-5.5	-7.0	0.5	1.32
HYDR	3.0	N/A	1.73	N/A

Because detailed velocity investigation is beyond the scope of this study, we adopt the velocities and JHD station corrections determined by Jupe et al. (1994) and also used by Phillips (2000) for the purpose of our comparative location analysis.

PREVIOUS RE-LOCATION WORK AT SOULTZ

Phillips (2000) relocated microearthquakes densely populating two volumes within the Soultz reservoir by manually re-picking phases. The "shallow" and "deep" clusters, contained roughly 200 and 350 events, respectively. Approximate locations of these clusters are indicated by gray circles on Figure 2.

Relocations from manual repicking collapsed the diffuse clouds of seismicity into narrow, intersecting features which exhibited temporal seismicity patterns showing initial tight clusters or alignments of events which subsequently spread into adjacent areas. These early alignments in the sequence have been interpreted as joints or joint/fracture intersections which represent fluid flow paths (Phillips, 2000).

Moriya et al. (2000) have addressed precise relative relocation of Soultz data through identification of highly similar earthquake doublets and multiplets. They have estimated relative hypocenter relocations using the cross-spectral method of Poupinet et al. (1984), and have observed numerous well-defined, narrow seismic lineations. Moriya et al (2000) further note that the high proportion of doublets and multiplets seen in the early period of the stimulation declines as time progresses. These observations are in good agreement with those of Phillips (2000) and Phillips et al. (2000), and we discuss them further, below.

AUTOMATIC RE-PICKING TECHNIQUE

We have developed a method which combines the most advantageous features of existing waveform cross-correlation-based phase adjustment methods (e.g., Dodge et al., 1995; Shearer, 1998; Rubin et al., 1998, 1999) in an automatic package which can be easily implemented even for very large (many thousands of events) data sets with a manageable amount of analyst interaction. Details of the signal processing, clustering and conjugate gradient techniques and our innovations are reported elsewhere (e.g., Aster and Rowe, 2000; Rowe et al., 2000); however, we summarize our method below.

Signal processing tools

Pick adjustments are estimated by comparing station-common waveforms having preliminary picks, and estimating relative lags between these picks. Lag estimation proceeds in two steps: a coarse integer-sample correlation, and a subsequent fine correlation step which provides a sub-sample relative lag refinement.

Coarse (integer sample) cross-correlation provides initial lag adjustments for each waveform pair. For multicomponent data, we apply polarization filtering, derived from eigenvalue decomposition of a joint signal covariance matrix, to improve P and S signal-to-noise levels prior to waveform comparison. We then calculate the cross-spectrum and cross-coherency for each pair of signals. The cross-coherency is used as a filtering function to adaptively downweight incoherent frequency bands prior to cross-correlation. We then calculate a suite of cross-correlation functions from overlapping, narrow-band-filtered representations of the (polarization- and coherency-filtered) waveform pair. The integer-rounded, energy-weighted mean correlation maximum and its associated standard deviation are calculated from these functions (Aster and Rowe, 2000; Rowe et al., 2000).

Waveform pairs exhibiting sufficiently high integer correlation values (> 0.8 for the Soultz catalogue) with low associated lag standard deviations (< 2 samples) are aligned to the nearest sample, rewindowed, and passed to the subsample lag estimation algorithm.

A cross-spectral phase slope method (e.g., Poupinet et al., 1984; Ito, 1985) is used to obtain subsample lag estimates following integer lag estimation. Because unbiased estimation of the cross-spectral phase slope is essential to obtaining accurate estimates of subsample lags and corresponding errors, we apply multitaper spectral estimation to this problem. A suite of eigenspectral measurements may be combined in a simple or weighted average to provide low spectral leakage, low variance spectral estimate and associated standard errors (e.g., Thomson, 1982; Park et al., 1987). Conjugate multiplication of corresponding eigenspectra for the waveform pair provides multiple linearly independent estimates of the cross-spectral phase, yielding a mean cross-spectral phase slope estimate with associated dimensionally meaningful standard deviation (e.g., Aster and Rowe, 2000, Rowe et al., 2000).

After integer and subsample lags and their respective standard deviation estimates are obtained for each waveform pair, we sum the lags to provide a total lag estimate; the standard deviation of the pick lag is the quadrature sum of the coarse and fine lag standard deviations (Aster and Rowe, 2000; Rowe et al., 2000).

Solving for consistent pick adjustments

The resulting system of inter-event pick lag constraints is solved using an iterative conjugate gradient approach formulated to operate efficiently with our sparse matrix storage scheme (Aster and Rowe, 2000; Rowe et al., 2000). The minimum L_1-norm residual solution is used (with a numerical modification for near-zero residuals to avoid

gradient discontinuities, e.g., Rowe et al., 2000) because of its improved robustness to outliers over L_2 methods (e.g., Parker and McNutt, 1980; Shearer, 1997). We further improve the solution by conservatively rejecting outliers in this massively overdetermined problem and successively re-solving the system (Rowe and Aster, 1999; Aster and Rowe, 2000; Rowe et al., 2000). We calculate 1-sigma error bars via Monte Carlo perturbation of the data vector.

Event association

We have found that, despite weighting the system constraints according to the associated lag standard deviations, solving the entire catalogue in one step does not provide optimal pick corrections because inconsistent waveform types are compared (Rowe et al., 2000). The choice of an appropriate event association method has proven to be crucial to improving the hypocenter locations. Because of the high degree of waveform variability within the small seismogenic volume, the Soultz catalogue has some corresponding difficulties which may not be encountered in other settings. Our event association and pick adjustment strategy is outlined below.

We have adopted an event clustering strategy based upon waveform similarity (e.g. Aster and Scott, 1993). We treat the maximum cross-correlation value for each event pair as a measure of distance in similarity space, so that the correlation matrix may be viewed as a spatial (dis-)similarity matrix.

We obtained good intra-cluster results when we selected one of the most variable phases and used it as our clustering basis. The S-wave arrivals for station 4550 exhibit tremendous variation in waveform, so we selected an extended (0.05 s) correlation window for that phase, filtered using the signal cross-coherency function for each

waveform pair. Cross-correlation of these waveforms provided a similarity matrix for a highly variable parameter in the seismogram database. We then applied a dendrogram-based hierarchical pair-group classification clustering scheme (e.g. Ludwig and Reynolds, 1988; Harris et al., 1999; Rowe et al., 2000) to build similarity groups. We have adopted a cross-correlation peak value of 0.75 as a minimum similarity criterion for any cluster.

Intra-cluster adjustments

Within each cluster we then cross-correlate all phases and solve the respective first-difference systems to determine optimal pick adjustments. Because of significant variation in initial pick error from event to event (adjustments range from a fraction of a sample to 50 samples or more), as well as variations in waveform signal to noise ratio, we find that correlation window length can strongly affect waveform alignment even within highly similar clusters. The best choice can vary from cluster to cluster, as well as from phase to phase within a cluster. We therefore calculate cross-correlations for a suite of window lengths for each phase, and choose the parameters which produce the best mean cross-correlation and lag standard deviation values for the phase being addressed (Rowe et al., 2000).

Once lags are estimated for all phases within a cluster, we solve for pick adjustments using our L1-norm-minimizing conjugate gradient solver, imposing a post priori zero mean constraint so that the mean of the adjusted picks corresponds to the preliminary pick mean. The corrected picks can provide very precise relative arrival times within a similarity cluster; however, inter-cluster relationships must still be addressed to achieve inter-cluster pick consistency, which may correspondingly correct

the inter-cluster spatial relationships. Some means of correcting this bias without relying upon preliminary locations or traveltimes is desirable.

Inter-cluster adjustments

We have found that a hierarchical correlation approach using stacked seismograms for each cluster provides good adjustments for inter-cluster relationships. For each phase, we calculate vertical waveform stacks within each of the clusters, where stacks are aligned on the adjusted picks (see Figure 3). In this way, each cluster may be viewed as having a composite catalogue location represented by the hypocenter calculated from its waveform stacks. To adjust the stacked waveforms for each phase (hence correcting for inter-cluster pick biases and, ultimately, inter-cluster relative centroid location), we apply the cross-correlation and conjugate gradient solution to the stacks. Each stack is assigned a relative pick lag which would serve to align stacks (mean picks) for every correlated cluster. We apply the stack lag to all corresponding phases for member events within each cluster, providing a final adjustment for inter-cluster relationships (Figure 3). The tightly constrained intra-cluster relative relationships are thus preserved while the addition of the stack lag information adjusts inter-cluster centroids with respect to one another.

Final pick adjustments

There may still be a catalogue-wide absolute picking bias for some phases. For example, successively later picks for smaller events with lower signal-to-noise ratios - particularly for more distant stations whose arrivals are more emergent to begin with - may result in a net late pick bias. Addressing any such final absolute pick offset may be appropriately handled through adjusting to a master event, application of an autopicking

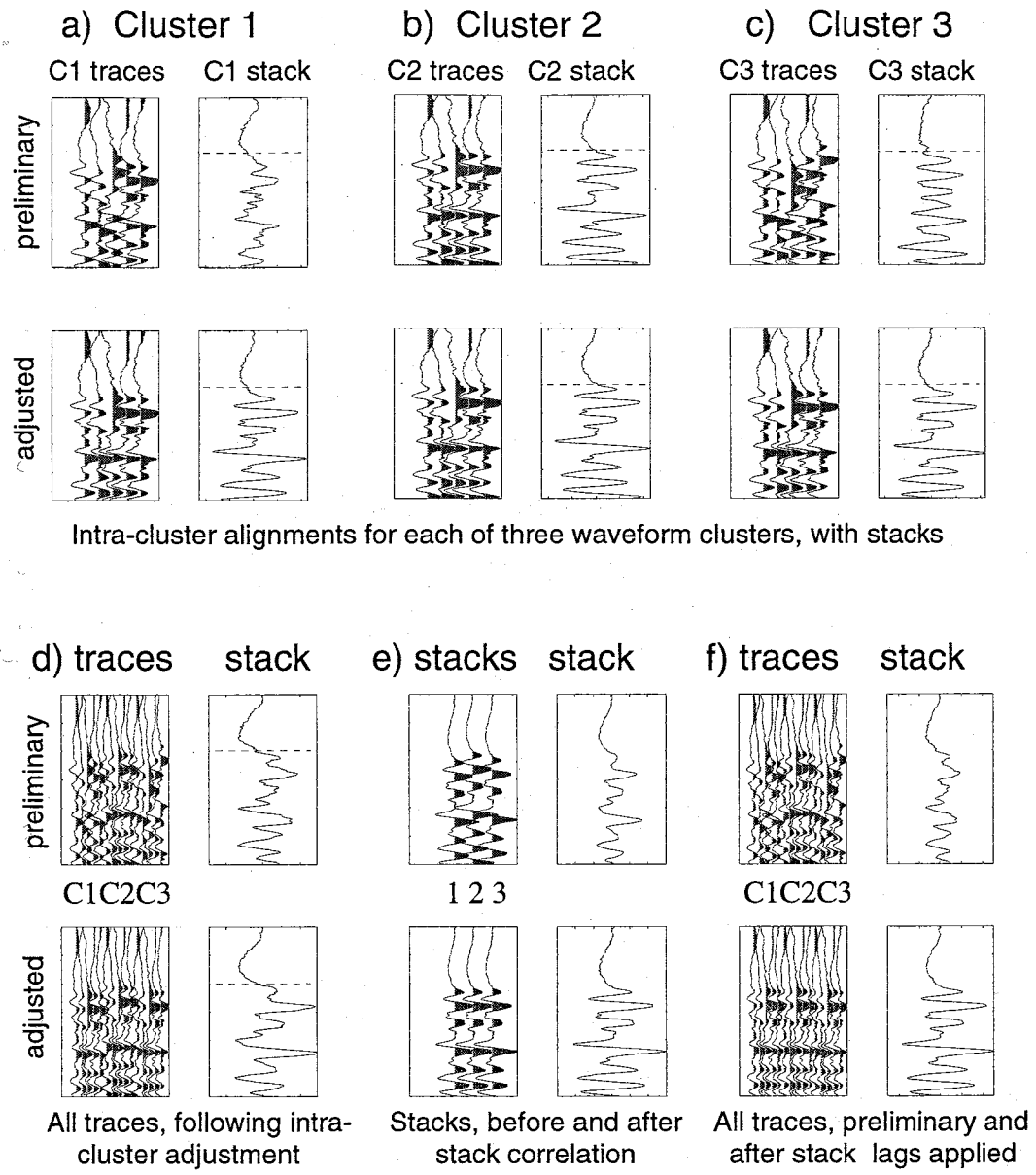


Figure 3: Hierarchical correlation, lag adjustment and stacking method. Panels (a)-(c) illustrate three synthetic clusters of five events each. Upper panels show traces aligned on preliminary picks and associated waveform stack; lower panels show traces aligned on adjusted picks, with associated stack. Horizontal dashed lines indicate pick times on the stacked trace. d) Clusters from a), b) and c) displayed together to show initial pick inconsistency (upper panel) and remaining biases among aligned clusters (lower panel). 3) Stacks from clusters of a) b) and c) showing stack mean pick misalignment (top) and aligned stacks (bottom). f) the same three clusters showing initial misalignments (top) and final, corrected alignments (bottom) after both intra-cluster (a-c) and inter-cluster (e) lags have been applied to the individual traces. From Rowe et al. (2000a).

routine to the second-order stacks, or application of consistent station corrections as derived through JHD or other joint location methods, although this latter approach does not separate velocity delays from pick error delays. We discuss this further below.

AUTOMATIC REPICKING TEST CASE

The Soultz reservoir seismicity presents an excellent test case for development and application of the automated repicking algorithm, although it also presents some significant relocation issues not encountered with many larger seismic networks in tectonically active regions. The primary difficulty lies in the small number of recording stations, as discussed above. Generally, relocation studies using waveforms from large seismic networks (e.g., Shearer, 1997, 1998; Rubin, 1998; Rubin, 1999) are able to neglect poor quality arrivals without compromising overall ability to obtain robust hypocenter locations. At the Soultz reservoir there are only four sites where sensors were operating during the 1993 stimulation; hence, discarding any phase arrivals can destabilize the hypocenter solution process, particularly given the difficult network geometry (Figure 2).

On the other hand, the sensors are situated in boreholes at depths between 1376 m and 2075 m, resulting in very low background noise and good microearthquake signals. Three of the instruments are multi-component sensors, which allow polarization filtering for improving signal correlation (particularly for S-waves).

We first tested our method using the deep cluster events of Phillips (2000) as a benchmark against which to compare our results (Rowe et al., 1998, Rowe and Aster, 1999; Aster and Rowe, 2000). The painstaking Phillips repicks and relocations were assumed to be the "correct" answer; step-wise comparison with these allowed us to gauge improvements in the development of our technique.

Initial correlation on the S-wave picks for station 4550 using a window length of 50 ms (about 5 cycles) provided a similarity matrix for clustering the 311 events within the Phillips deep cluster which were included in our catalogue. This resulted in eighty similarity groups of varying sizes, dominated by doublets (two-event clusters). Figure 4 shows a histogram of similarity cluster membership for this trial dataset. Eighteen events were orphaned (unassociated) in the clustering process; the remaining 293 were addressed in the analysis.

Applying the automatic repicking algorithm to P and S phases for the deep cluster, we demonstrate the resulting waveform alignment in Figure 5. The uppermost panels for each phase illustrate alignment on preliminary analyst picks. Center panels represent alignments after intra-cluster adjustments have been made; the lowermost panels show alignment following correction for inter-cluster consistency. Stacked waveforms for each alignment plot are shown on the right, with the aligned picks indicated by horizontal black bars. The increase in stack amplitude provides a qualitative measure of the repicking success.

Note that Figure 5 does not show the filtered and projected waveforms which were cross-correlated to obtain the pick lags. Because the filtering and optimal signal rotation may be different for every inter-event cross-correlation, it is not possible to show these in a single figure. We show instead alignments of unfiltered traces. For P-wave alignments, we show the vertical component, and for S-wave alignments, we display the east-west (X) horizontal component. Improvement in waveform alignment is especially marked for S-waves. For instance, at station 4616 the summed S-wave stack amplitude after repicking is nearly three times that of the preliminary stack.

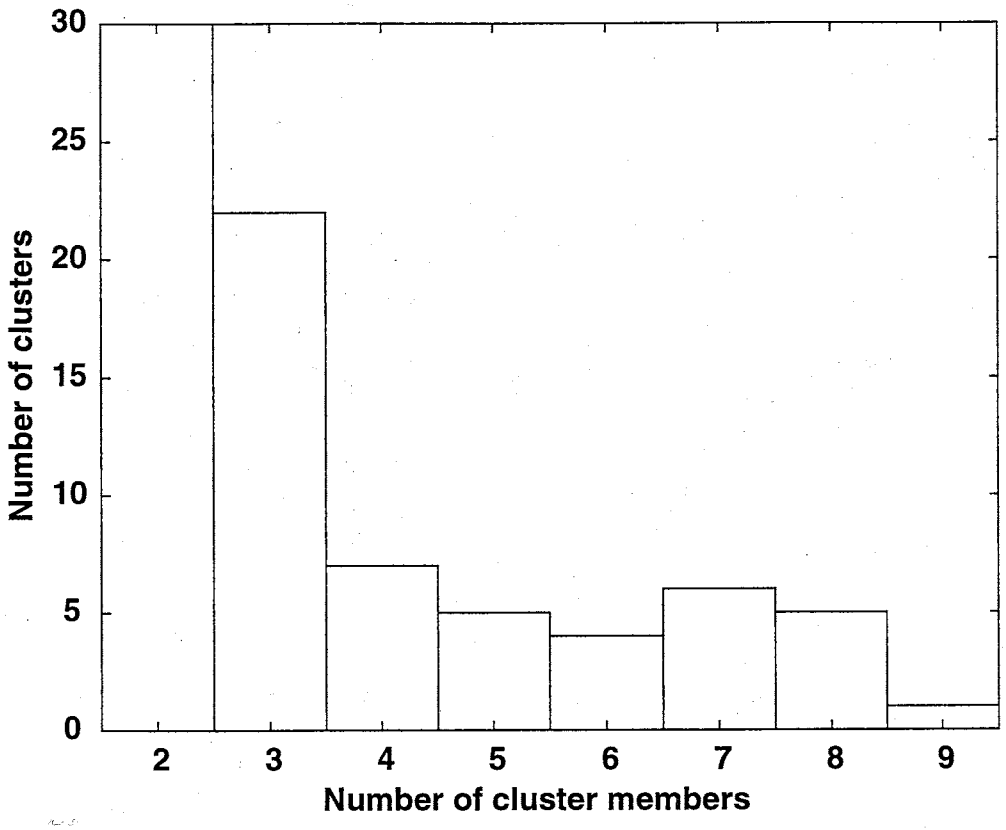


Figure 4: Earthquake cluster sizes for Phillips' deep cluster. Histogram shows the number of families for each category of doublet, triplet, quadruplet, and so forth. A total of eighty clusters was identified.

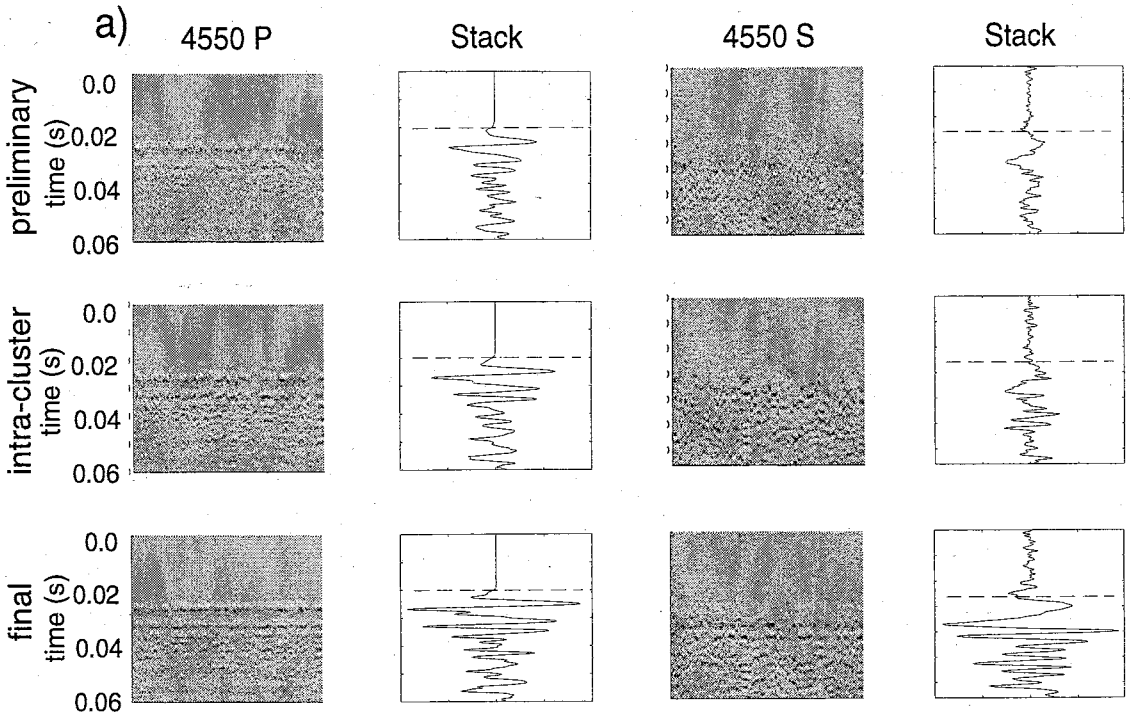
Figure 5: Waveform alignment plots for events in the deep cluster (Phillips, 2000). Waveforms are plotted as a grayscale surface with white peaks and dark troughs.Each phase is depicted in six panels. Waveform alignments are on the left, stacks of the waveforms are on the right. Top panels show waveforms aligned on preliminary picks. Center panels show waveforms (and stack) aligned on the repicks arising from intra-cluster relative repicks. Lower panels show alignments and stacks of waveforms after inter-cluster adjustments have been made.

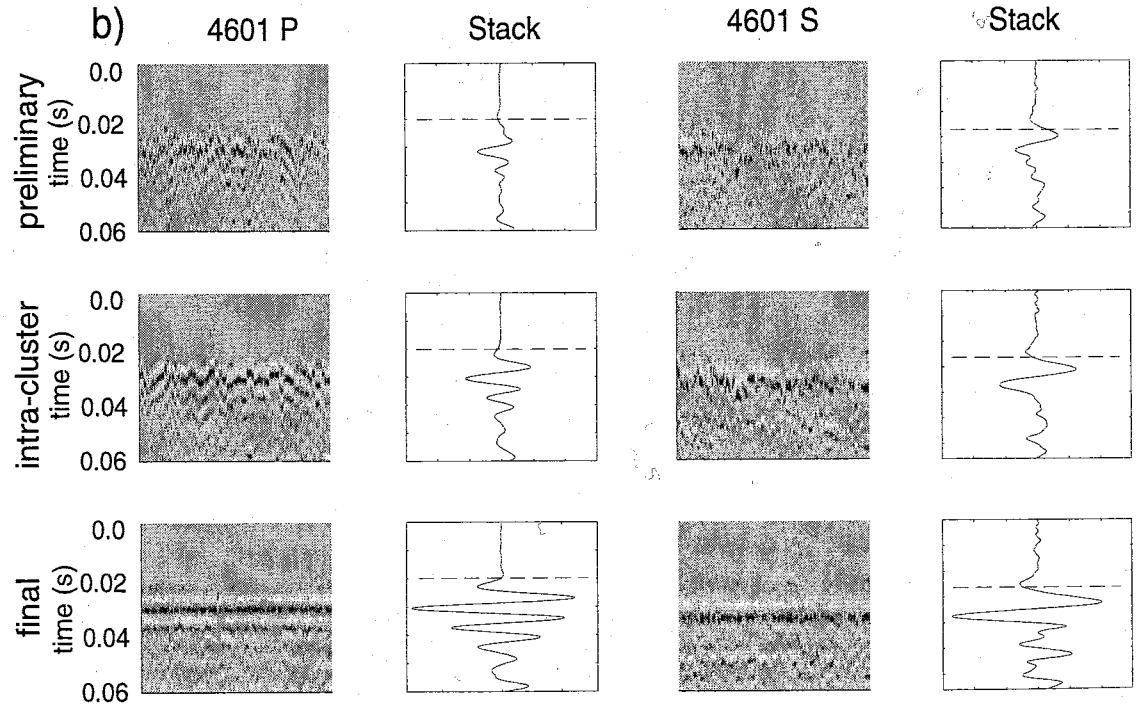
a) P-waves for station 4550 are shown in the two left-hand columns. S-waves for station 4550 are shown in the two right-hand columns.

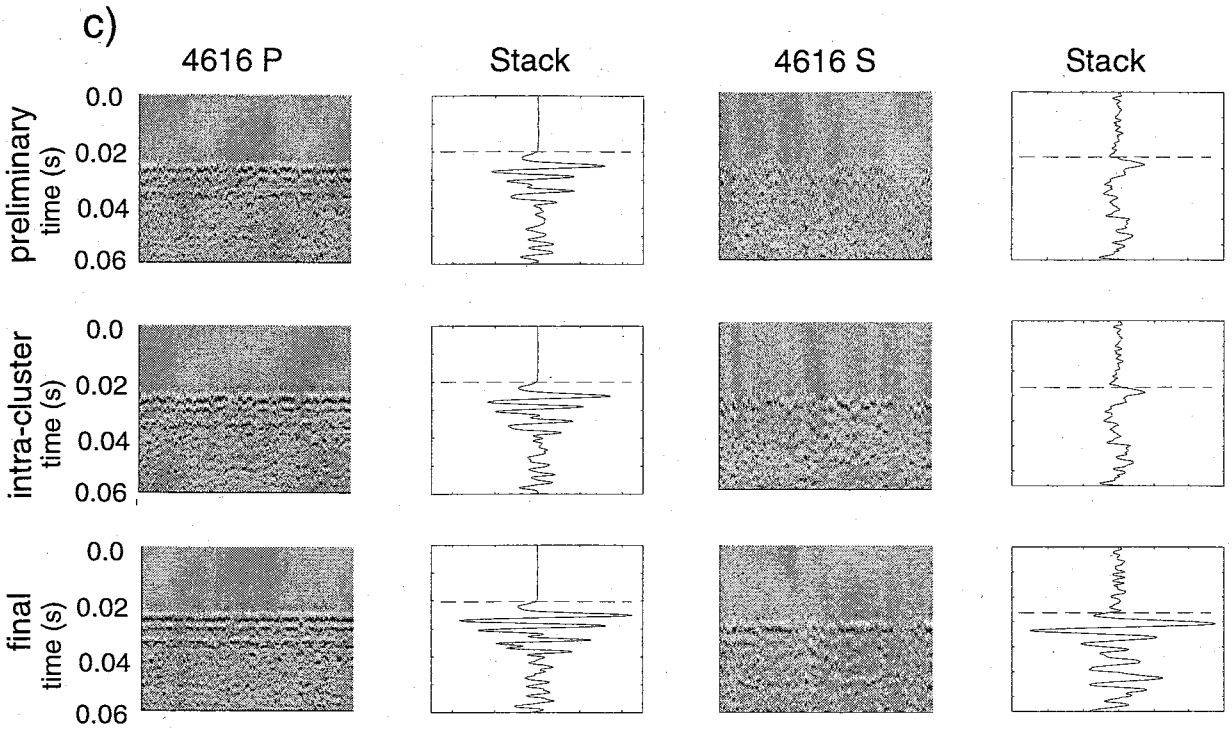
b) P-waves for station 4601 are shown in the left-hand columns; S-waves are shown in the two right-hand columns.

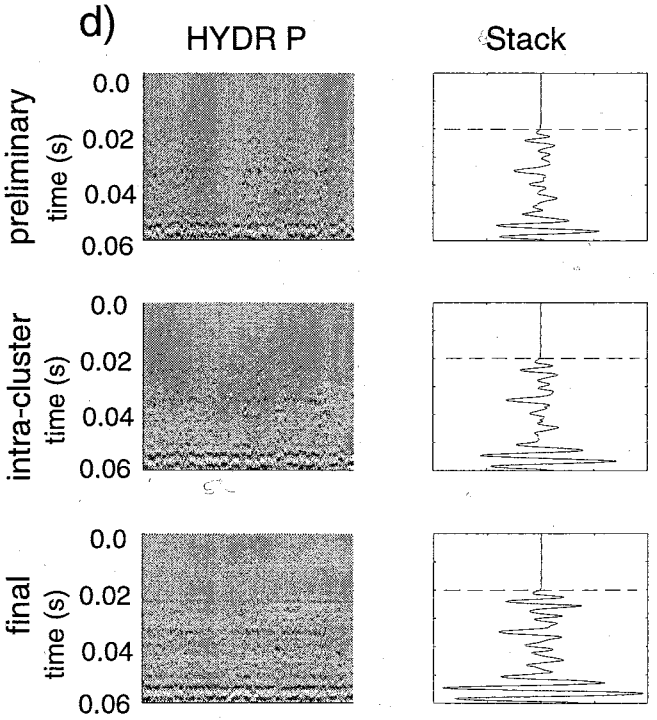
c) P-waves for station 4616 shown on the left-hand side, S-waves for station 4616 shown on the right.

d) P-waves for the hydrophone. No preliminary S was picked for the hydrophone, as this instrument was underdamped and the waveforms were often clipped.


a)
4550 P
Stack
4550 S
Stack
preliminary
intra-cluster
final
time (s)
0.0
0.02
0.04
0.06


b)
4601 P
Stack
4601 S
Stack
preliminary
intra-cluster
final
time (s)
0.0
0.02
0.04
0.06


c)
4616 P
Stack
4616 S
Stack
preliminary
intra-cluster
final
time (s)
0.0
0.02
0.04
0.06


d)
HYDR P
Stack
preliminary
intra-cluster
final
time (s)
0.0
0.02
0.04
0.06

After cross-correlating and solving for consistent pick adjustments, we re-located the deep cluster earthquakes using the same location algorithm (single event, damped least squares) and velocity parameters (Vp= 5.85 km/s, Vs=3.34 km/s) determined by Jupe et al. (1994) and used by Phillips (2000). Figures 6a and 6b illustrate hypocenter locations for preliminary picks and manual re-picks, respectively. In Figure 6c we show hypocenters calculated using automatic intra-cluster repicks, to illustrate the increased degree of clustering obtained by this step. These hypocenters correspond to the waveform alignments in the center panels for each phase shown in Figure 5. Tight intra-cluster alignments result in increased clustering among hypocenters, but the large inter-cluster biases have not yet been addressed. Note that the structures in Figure 6b are still not clear in Figure 6c, although hypocenters exhibit significantly increased clustering over the preliminary locations of Figure 6a. Figure 6d shows the automatic repicks following inter-cluster stack correlation and adjustment (corresponding to picks aligned in the lowermost panel for each phase in Figure 5). The clearly defined, near-orthogonal intersecting joint features illuminated by the manual repicking in Figure 6b and discussed in detail in Phillips (2000) and Phillips et al. (2000) are also clearly distinguishable in Figure 6d. Median hypocenter mislocation, compared to relocations using the Phillips picks, is reduced from 31 to 7 m after we have adjusted centroid median biases.

Because Phillips correlated peaks and troughs with the visual cross-correlation, then applied station corrections to his picks to approximate first-break consistency for locating events, it is difficult to compare directly the relocation quality in terms of hypocenter RMS traveltime errors; however, we iteratively adjusted the Jupe et al. (1994) JHD station corrections to minimize mean traveltime residuals for automatic repicks and pre-

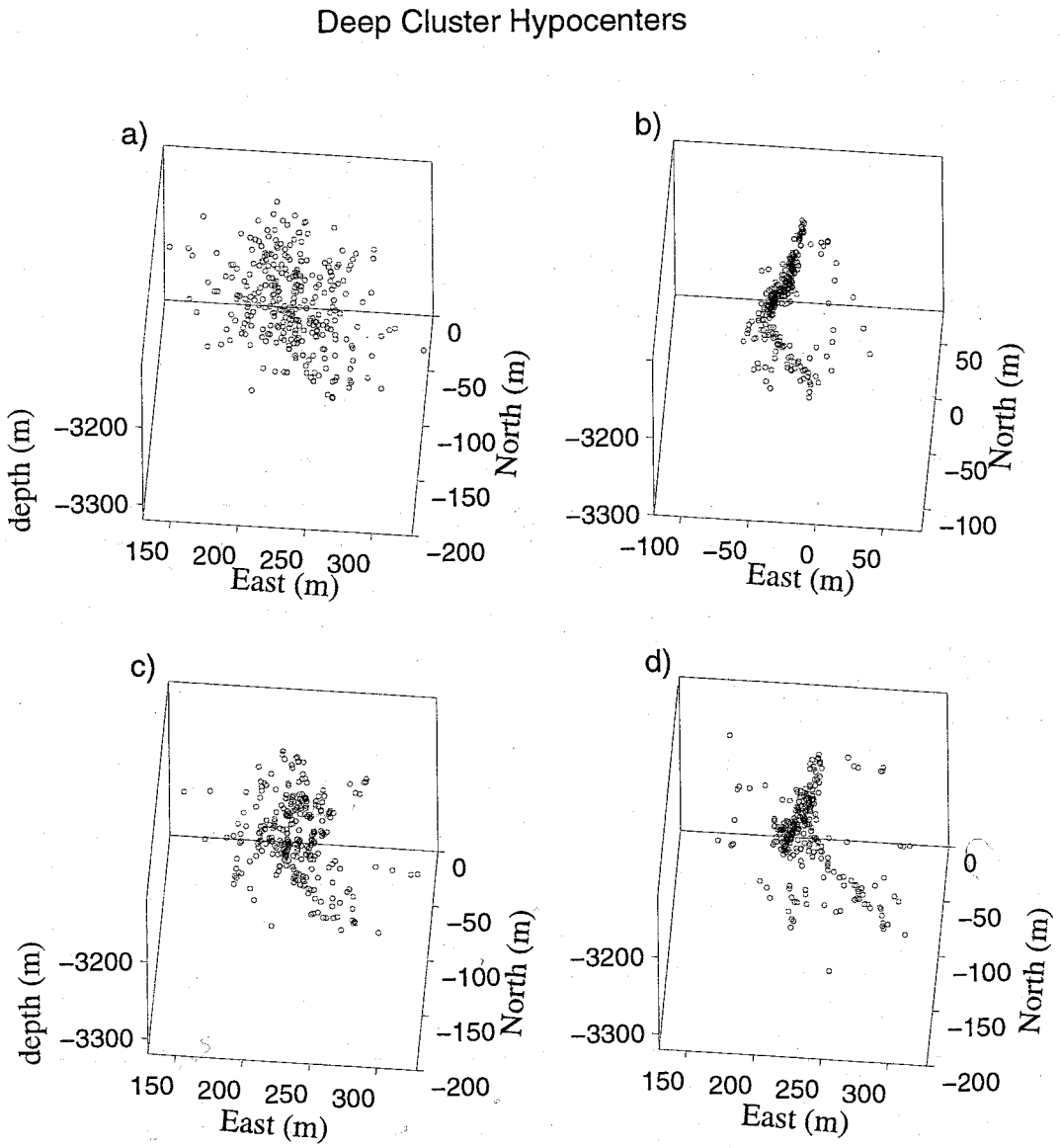


Figure 6: Hypocenter locations for Phillips' deep cluster. a) Preliminary locations, based on initial, analyst picks. Note the random scatter in the distribution of seismicity. b) Relocations based on Phillips (2000) manual repicking. The diffuse cloud of hypocenters (a) has converged into two nearly orthogonal, intersecting joint features and a few scattered events. c) Relocations based on intra-cluster pick adjustments (corresponding to waveform alignments in center panels of Figure 5). d) Relocations after inter-cluster stack adjustments have been applied to individual traces (corresponds to alignments in lower panels of Figure 5). Although Figure 6b exhibits a centroid shift from 6a,c and d, all four panels are plotted at equal scales.

liminary picks. The resulting RMS traveltimes errors are 0.68 ms, 0.25 ms and 0.14 ms for preliminary, automatic and manual picks, respectively, with associated standard deviations of 0.62 ms, 0.34 ms and 0.24 ms, respectively.

APPLICATION TO THE LARGER DATA SET

When we had completed testing of the automatic algorithm using the Phillips (2000) deep cluster, we turned to the task of relocating the ~7,000 events in the Soultz catalogue that had seven preliminary phase picks.

Relocations

Processing the ~7000-event catalogue proceeded in the hierarchical correlation/ clustering method as discussed for the deep cluster (above). Following cross-correlation on the S-waves at station 4550, the algorithm segregated the catalogue into 2408 clusters, plus 661 unassociated (orphaned) events. Figure 7 shows the resulting cluster memberships. We show the number of clusters of each size (from 2 to 18 members) in Figure 7a, while plotting the cumulative proportion of the catalogue comprised of the different cluster sizes in Figure 7b. Picking inconsistencies were resolved within similarity clusters with median lag adjustments ranging from 0.251 ms for station HYDR P-picks to 1.6 ms for station 4601 S picks. Intra-cluster adjustments are detailed in Table 2.

Table 2: Median intra-cluster relative pick adjustments and standard deviations

Station	median P lag (ms)	P st.dev (ms)	median S lag	S st.dev (ms)
4550	0.34	0.03	0.76	0.05
4601	1.1	0.06	1.6	0.06
4616	0.38	0.03	0.8	0.06
HYDR	0.25	0.03	N/A	N/A

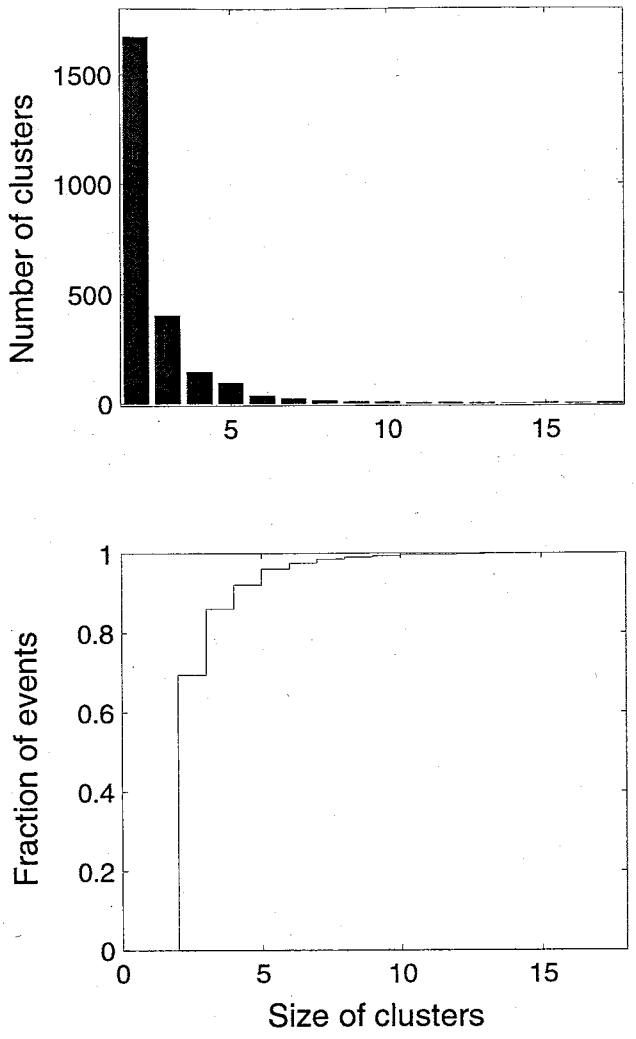


Figure 7: Cluster memberships for the ~7000-event data set. a) Histogram shows the number of families for each category of doublet, triplet, quadruplet, and so forth. A total of eighty clusters were identified. b) Cumulative fraction of the catalogue assigned to clusters of two, three, four and so forth. Note that doublets and triplets comprise ~87% of the data, indicating that adjusting relative pick lags among clusters is important to the final resolution of spatial relationships among earthquakes.

After adjusting each of the seven phase picks for consistency within each of the 2408 clusters, the aligned waveforms for each phase were stacked to provide a composite waveform (e.g., Figure 3). These cluster stacks were then correlated with corresponding stacked waveforms for all other clusters so that inter-cluster picking inconsistencies might be resolved, and the final pick lag adjustments made. Median inter-cluster stacked waveform adjustments are shown in Table 3. Maximum inter-cluster lag adjustments were 17.2 ms and 11.1 ms for S waves and P waves, respectively; both maxima were observed for station 4601, the most distant station from the seismogenic volume (Figure 2).

Table 3: Median inter-cluster stack pick adjustments and standard deviations

Station	median P lag (ms)	P st.dev (ms)	median S lag (ms)	S st.dev (ms)
4550	0.77	1.2	1.5	2.4
4601	1.8	2.7	3.0	4.7
4616	0,85	1.3	2.0	2.8
HYDR	0.55	0.96	N/A	N/A

Automatic repicking and relocation of our ~7000 event dataset yielded significant improvement in illumination of interpretable seismogenic features. Median RMS travel-time error for repicked hypocenters decreased from 3.741 ms to 3.739 ms compared to original picks, using the station corrections derived from JHD (e.g., Jupe et al., 1994). This moderate change of 0.2% indicates that overall location errors are far more heavily influenced by the oversimplified velocity model and the station correction parameters, which were derived to optimize fitting of the preliminary picks for a larger dataset, than by moderate pick adjustments; however, waveform alignments and improved resolution of fracture and joint features indicate that the pick adjustments are providing improved hypo-

center locations. Figures 8-11 show the results of relocating our ~7000-event dataset. We show map-view and three-dimensional projections for depth slices at 2500-2900 m (Figure 8), 2900-3100 m (Figure 9), 3100-3300 m (Figure 10) and 3300-3700 m (Figure 11). The three-dimensional projections have been chosen for optimal view angle to show the emerging structures within the seismicity cloud at each level. In each of Figures 8-11, we depict preliminary hypocenter locations in the left-hand panels, and relocated hypocenters in the right-hand panels. We have used the same SEL location algorithm (Phillips, 2000) for both preliminary and repicked events to demonstrate the level of improvement which is possible solely from automatic repicking.

Relocated shallow events are shown in Figure 8. Figure 8a shows these events in map view. Black arrows (B) and (C) indicate view azimuths in Figures 8b and 8c. Shallow earthquakes exhibit roughly parallel lineations in map view, striking N40°W in the more southerly part of the reservoir, although the seismicity does not define these trends as clearly in the more northerly part. This orientation is similar to the orientation of one of the populations of hydrothermally sealed fractures identified by Genter and Traineau (1996). To the north, lineations suggest fractures or faults with an orientation of approximately N20°E (indicated by gray arrows I, II at the top of the figure). An elevated view looking S50°E with a 55° inclination (Figure 8b) suggests that the N40°W seismicity trend dominating the southern part of the reservoir at these depths is made up of discrete, subparallel planes (indicated by gray arrows III, IV, V) steeply dipping towards the northeast. These structures are not discernible at all in the preliminary hypocenter cloud (8b, left-hand panel). In Figure 8c we show a cross-section looking towards N55°W. This shows

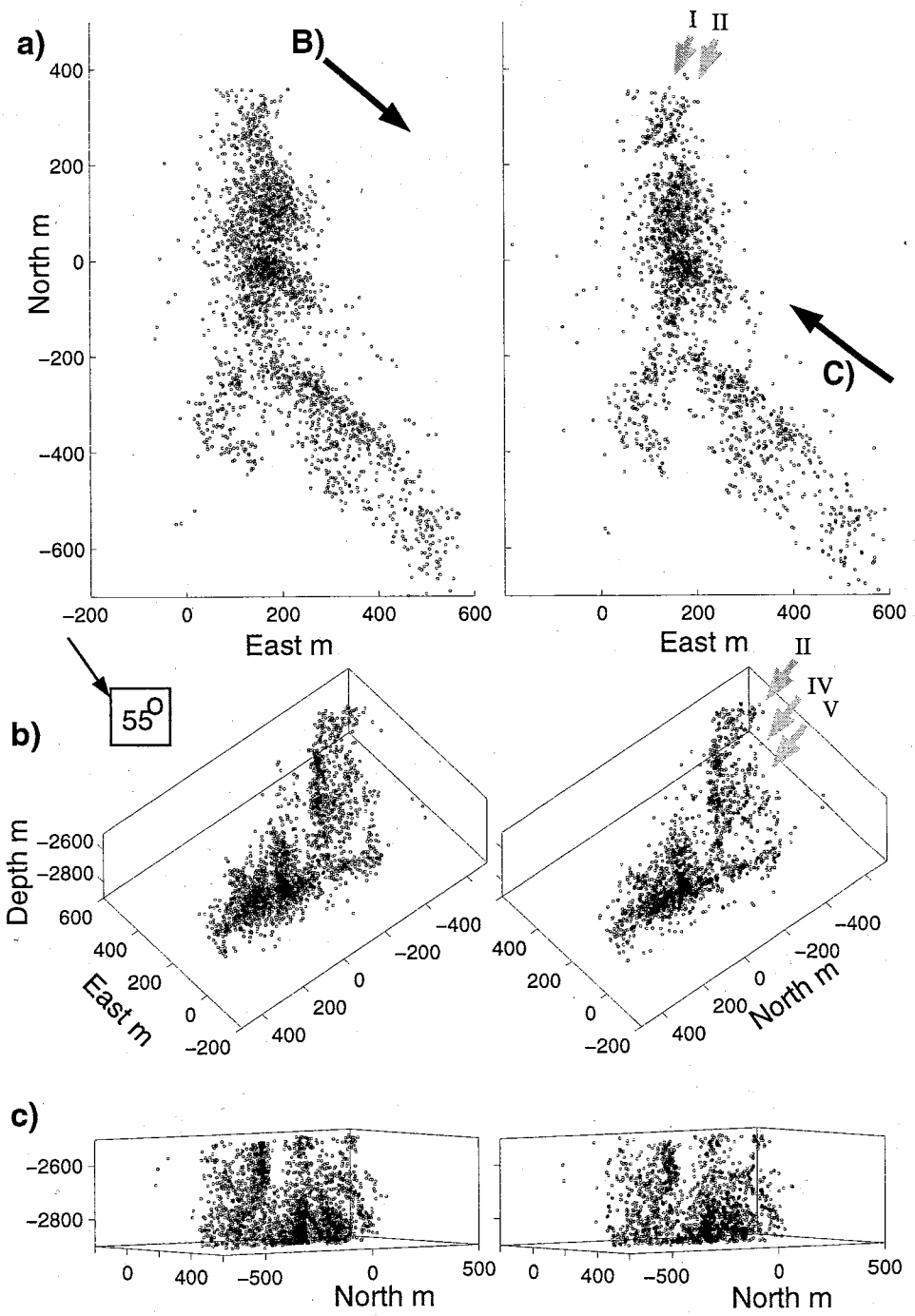


Figure 8: Preliminary (left) and relocated (right) hypocenters for shallow (2500 m to 2900 m depth) Soultz seismicity. Details are discussed in the text. a) Map view. Arrows (B) and (C) indicate the view directions for Figures 17b and 17c. b) Steeply inclined view at 55° of shallow seismicity, looking towards S50°E. c) Cross-section looking N40°W along strike of the segmented, southern portion of the shallow reservoir.

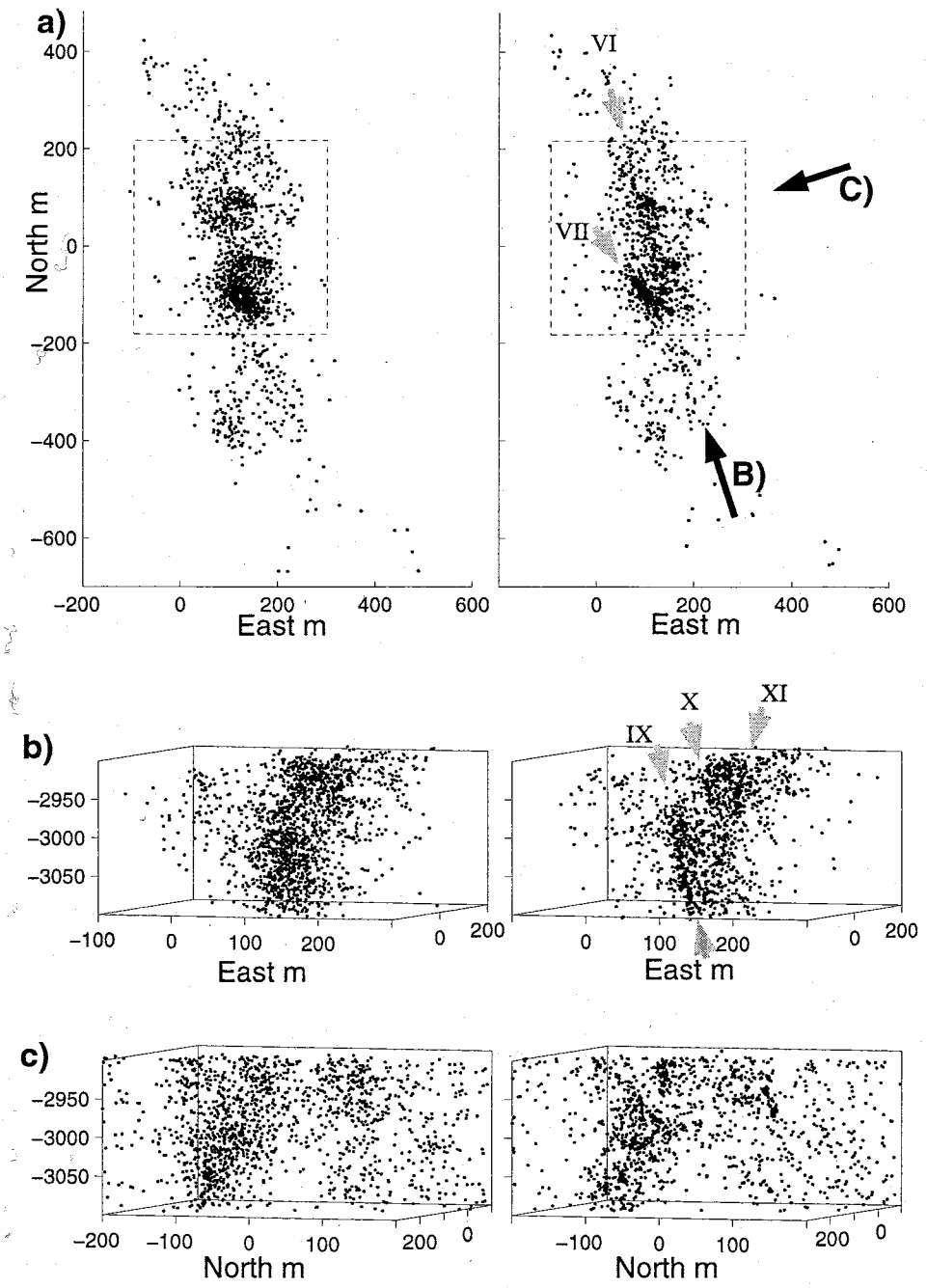


Figure 9: Intermediate depth seismicity (between 2900 and 3100 m). Details are discussed in the text.a) Map view. Cross-section view angles in (b) and (c) are indicated by direction arrows (B) and (C). b) Cross-sectional view looking N18W, along strike of the clustered features seen in map view. c) Broadside view looking S72°W.

the steeply eastward-dipping nature of the southern lineations, as well as a northerly dip for the isolated planes of seismicity seen in 10a at 200-300 m north.

Intermediate-depth events (between 2900 and 3100 m) are plotted in Figure 9. In map view the northwesterly-striking features may still be seen, although lineations are less laterally extensive. Two linear features are suggested, one striking N15°W (gray arrow VI) and one approximately N30°W (gray arrow VII). Seismicity did not expand as far to either side of the borehole at this level as it did at a shallower depth, suggesting a change in the fracture system and its connectedness at approximately 2900 m depth. The dashed boxes in Figure 11a indicate the area displayed in close-up cross-section in 9b and 9c. Black arrows (B) and (C) indicate view azimuths for Figures 9b and 9c. In Figure 9b, looking N10°W, we see that after relocation the hypocenters define planar features which dip steeply to the north (gray arrows IX and X), but also a lineation suggesting an intersecting feature (gray arrow XI). Viewing these features broadside in Figure 9c (perpendicular to strike) illuminates small linear seismicity clusters within their planar surfaces.

Deeper (between 3100 and 3300 m) events are shown in Figure 10. Relocated hypocenters exhibit considerable improvement in clustering as seen in map view (Figure 10a), particularly in the dense concentration of events at (100,-50). These events correspond to the depth of the Phillips (2000) deep cluster, discussed earlier. Black arrows (B) and (C) indicate view azimuths in Figures 10b and 10c. In Figure 10b we see the inclined fracture ("secondary") plane (gray arrows XI and XII) of the Phillips deep cluster (see Figure 6) as well as the northerly-trending ("primary") plane (gray arrows XIII and XIV) that it intersects. Relocated seismicity suggests additional lineations parallel to these major features which may indicate a series of subparallel seismogenic fractures activated

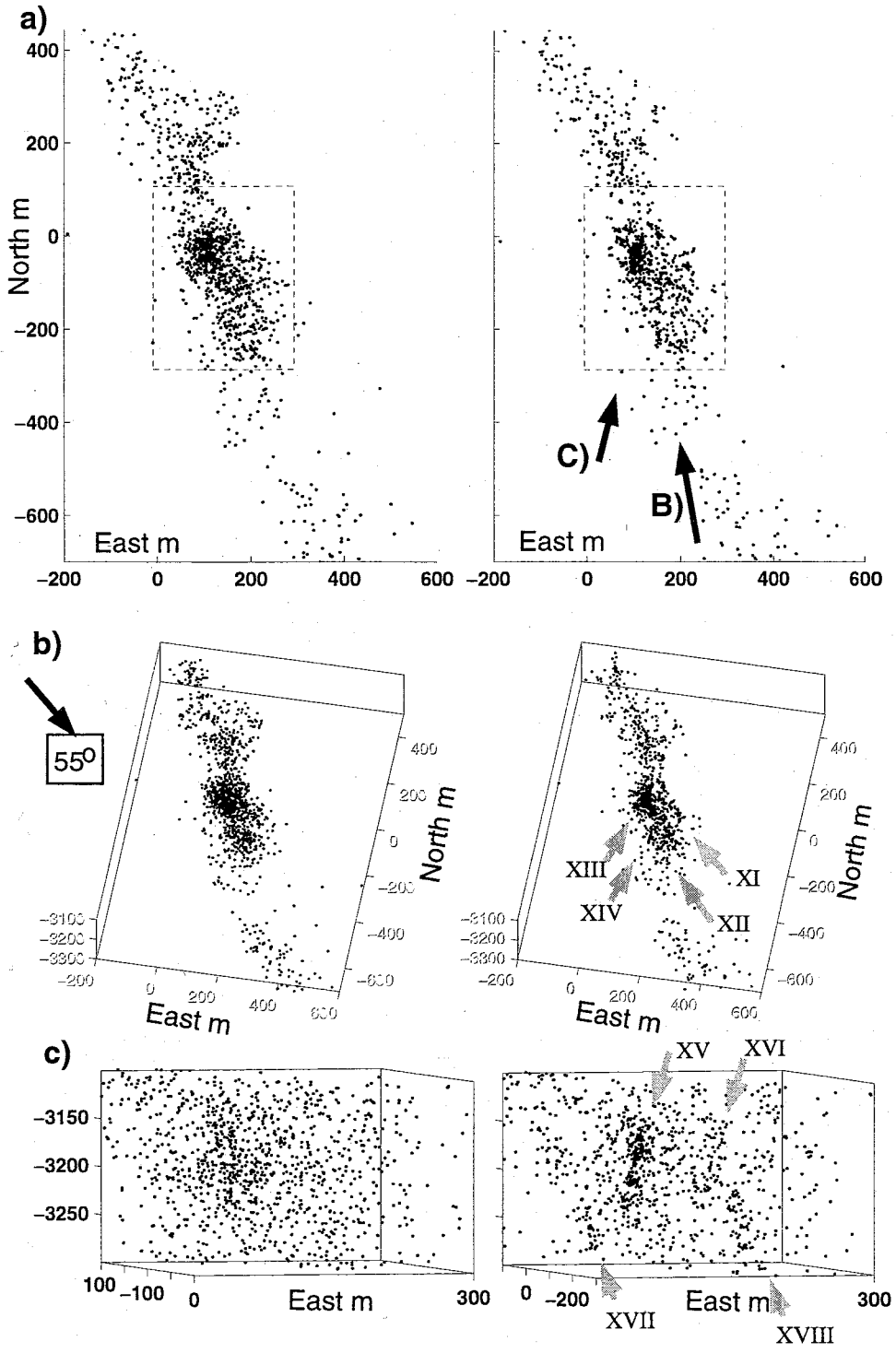


Figure 10: Deep seismicity between 3100 m and 3300 m. Details discussed in the text. a) Map view. Dashed box outlines the close-up elevation shown in c). View azimuths for (b) and (c) are indicated by arrows (B) and (C). b) Elevated view at 55° inclination, azimuth N10°W, approximately coincident with Phillips' deep cluster intersection axis. c) cross-section looking N14E, along strike of Phillips' "primary" plane.

by the reservoir stimulation. The cross-section in Figure 10c looks towards N14E, along strike of the Phillips "primary" plane. The steeply westward-dipping lineation (gray arrow XV) is echoed by another, less-well-defined plane some 100 m to the east (gray arrow XVI); two deeper lineations dipping steeply towards the east (gray arrows XVII and XVIII) are suggested in this view as well.

Figure 11 shows the deepest seismicity at Soultz associated with the 1993 injections, extending from 3300 m to 3700 m depth. In map view (Figure 11a) we see the diffuse cloud is resolved after relocation into a dense, northwesterly-trending feature that appears to be comprised of discrete patches of seismicity. Black arrows (B) and (C) indicate view azimuths for Figures 11b and 11c, respectively. In a cross-section looking N40°W (Figure 11b), this concentrated zone of seismicity is revealed to be composed of two sets of roughly planar features, striking N40°W. The shallower features, extending to ~3500 m depth, appears to dip at an angle of about 60° towards the south (gray arrow XIX), whereas the deeper feature is a well-defined plane extending from ~3500 m to ~3700 m depth, dipping steeply to the north (gray arrow XX). Viewed broadside (Figure 11c), both shallow and deep planes consist of discrete seismicity patches which may correspond to distinct rupture zones on the seismogenic planes. Quasi-linear edges (gray arrow XXI) to these zones suggest truncation by some linear or planar feature intersecting the seismic plane.

DISCUSSION

The microearthquakes comprising this dataset are the result of two different periods of reservoir stimulation, in September and October, 1993. Figure 12 illustrates preliminary hypocenter locations plotted as a function of time for the September stimulation. In

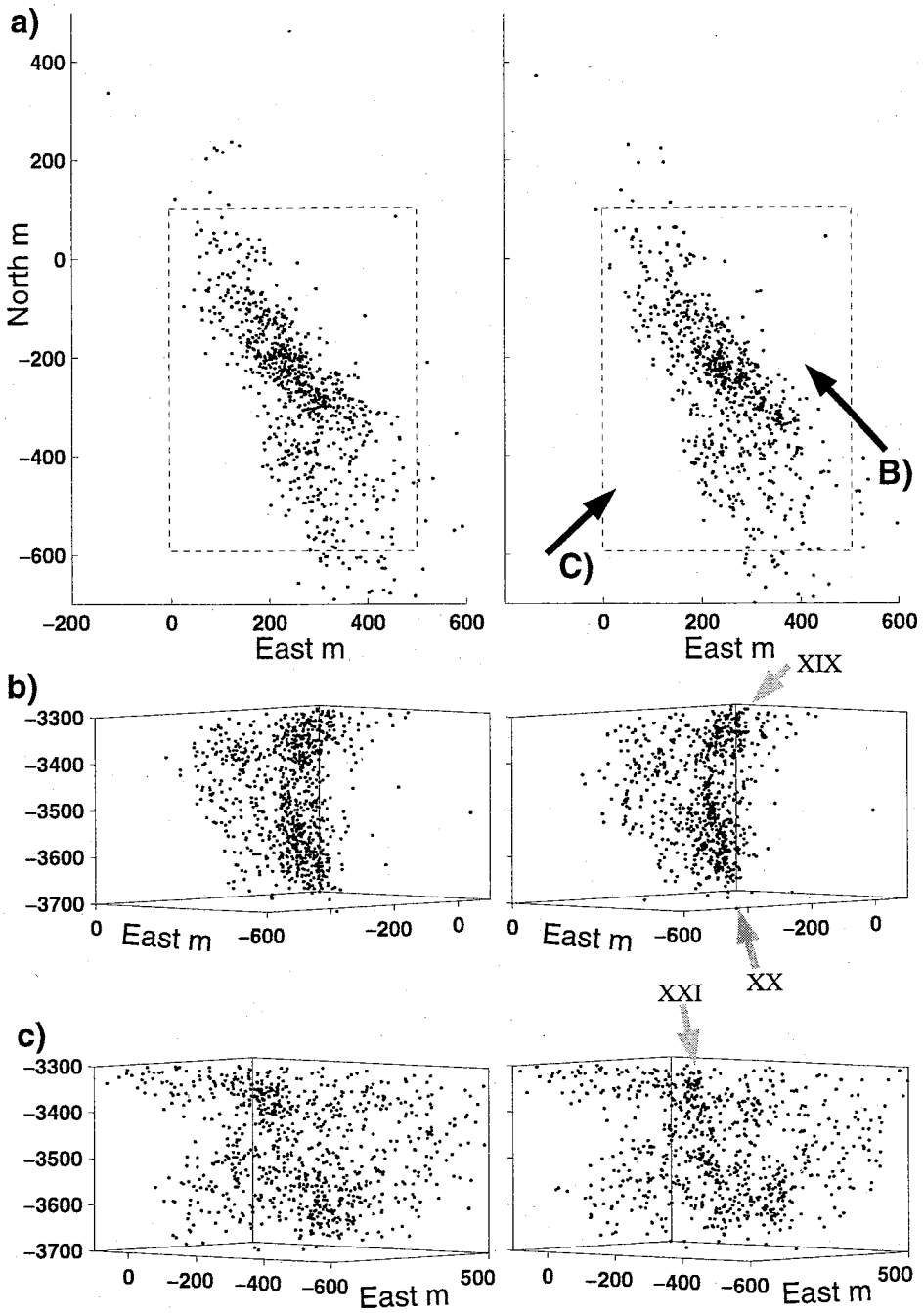


Figure 11: Deep seismicity (3300-3700 m) associated with 1993 injection. Details are discussed in the text. a) Map view. Dashed box represents region shown in the elevations in b) and c), whose azimuths are indicated by arrows (B) and (C). b) view looking N40°W along strike of the general trend in (a). c) Broadside view, perpendicular to (b).

Figure 12a we show hourly event counts (for the 7,000 microearthquakes we studied). Figures 12b-12d illustrate the seismicity distribution through time in East-West (X), North-South (Y) and depth (Z) dimensions, respectively. Approximately 80% of the microearthquakes occurred during the September stimulation, during which ~45,000 m^3 of water were injected over an open-hole depth range between 2800 and 3500 m at rates of up to 50 l/s and overpressures of ~10 MPa (Jupe *et al.*, 1998). The bottom portion of the injection well GPK1 was filled with sand (depths > 3480 m) to prevent flow into a previously identified, deep fracture zone. Seismicity during this first stimulation began at a depth of ~2900 m, adjacent to the injection site, and expanded over time in a roughly ellipsoidal cloud oriented NW-SE (Jupe *et al.*, 1998).

Histograms of hypocentral depths of the seismicity resulting from the two stimulations are shown in Figure 13. The September stimulation, shown in Figure 13a, exhibits a strong peak in seismicity at a depth of ~2900 m. The October stimulation has the greatest number of its events occurring at ~3600 m depth (Figure 13b). The time-varying hypocentral depth plot in Figure 13d shows this, and demonstrates that the removal of sand from the final 110 m of well GPK1 prior to the second stimulation causes the primary seismic activity to be generated by new slip within the deep fracture system, while the shallower portion remains comparatively quiet until overpressures are raised sufficiently to induce further failure on the previously de-stressed fracture system (Jupe et al., 1998).

The greatest concentration of events within a volume might be expected to yield the greatest number of similar earthquakes. Intuitively we can look at the growing volume of seismicity and suspect that in the earlier hours of stimulation, when seismicity is concentrated within a small volume, many similar earthquakes should be observed. Moriya et

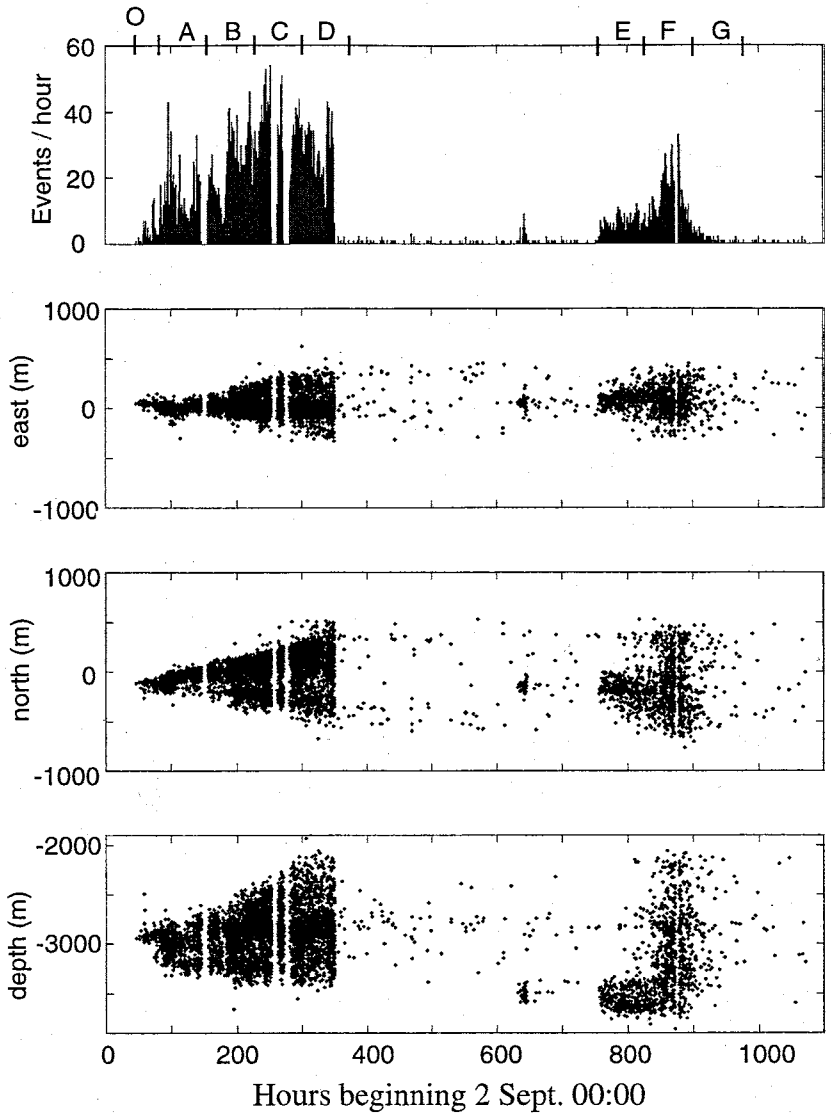


Figure 12: Spatial development of Soultz seismicity as a function of time, during the September and October 1993 injections over open-hole depths of 2800-3500 m in well GPK1 (Figure 2). a) Hourly event counts. Segments a-d and e-g represent 72-hour time intervals examined in Figures 14 and 15. b) Longitude (meters east). c) Latitude (meters north). d) Hypocentral depth.

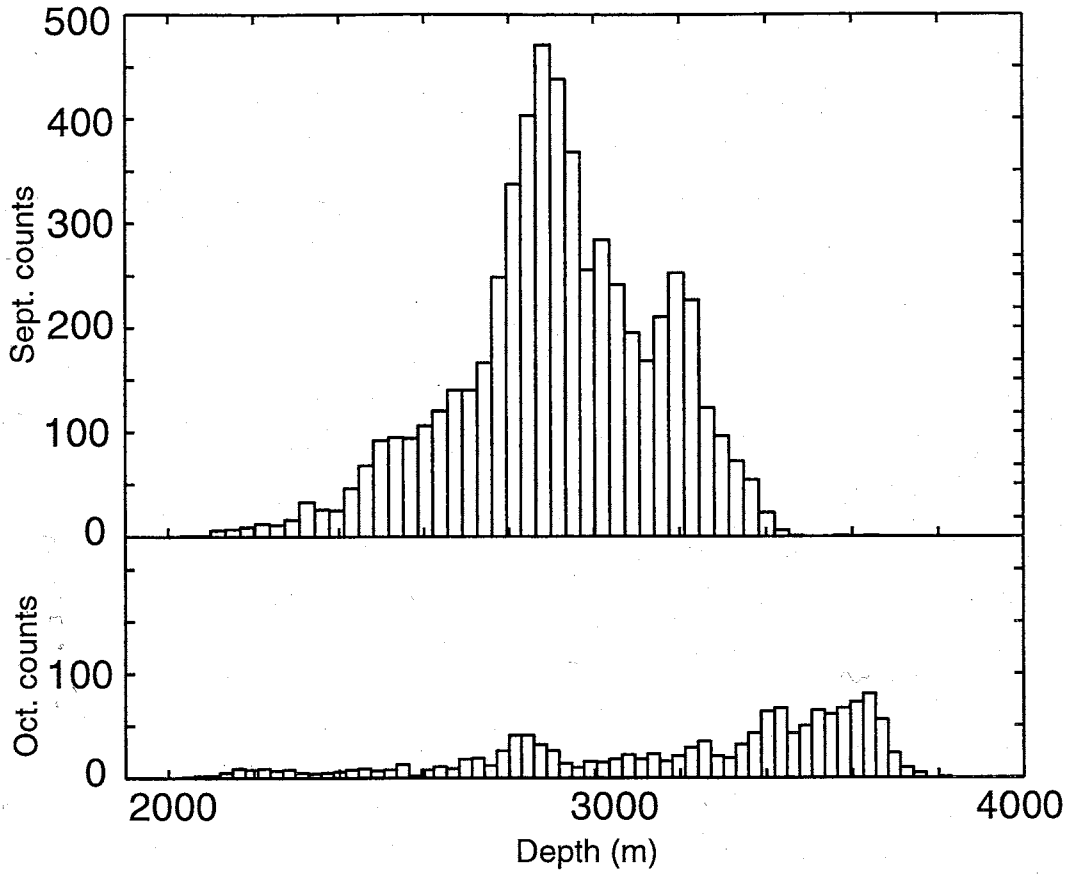


Figure 13: Histograms of hypocentral depths during the September and October, 1993 reservoir stimulations (Figure 12) is dominated by earthquakes occurring at about 2900 m depth. After removal of the sand at the bottom of GPK1, the October stimulation (b) exhibits greatest activity at 3400-3600 m depths, with a secondary peak at the fracture zone near 2900 m depth

al. (2000) have shown through their analysis of doublets and multiplets that seismicity is dominated by doublets/multiplets during early stimulation, whereas the later activity has a significantly lower proportion of highly similar events. In his manual re-analysis of two seismicity clusters, Phillips (2000) noted that tight lineations of relocated hypocenters are formed prior to the spreading of seismicity into adjacent planar features; he interpreted the tight lineations of hypocenters as primary flowpaths for the injected fluid. The observations of Moriya et al. (2000) and Phillips (2000) are complementary and concordant. Waveform similarity (*i.e.* membership in multiplets) is more likely to be observed within the early, tightly clustered events where source and path effects may be consistent.

We expand upon these observations by calculating time-varying and spatially varying hypocenter density within the stimulated volume. The seismogenic volume in any time interval $d\tau$ may be estimated as the product of the maximum X, Y and Z dimensions during $d\tau$. These dimensions are determined to first order by estimating a three-dimensional envelope about the time-varying seismicity. Figure 14 shows event counts, approximate reservoir volume and estimated mean event density for the first five intervals (from Figure 12) following the onset of stimulation. Event count per unit volume appears to decline in a roughly logarithmic fashion (Figure 14c). The highest value of event density, $10^5/km^3$, corresponds to the Moriya et al. (2000) estimated 90% doublet/multiplet fraction for events. The lowest density value, ~$1400/km^3$, corresponds to Moriya et al.'s approximately 50% multiplet fraction.

Identification of planar and linear structures which may indicate likely fluid flow paths is one of the primary goals of any geothermal reservoir-related evaluation of microseismicity. Fehler et al. (1987) developed an innovative method of identifying likely

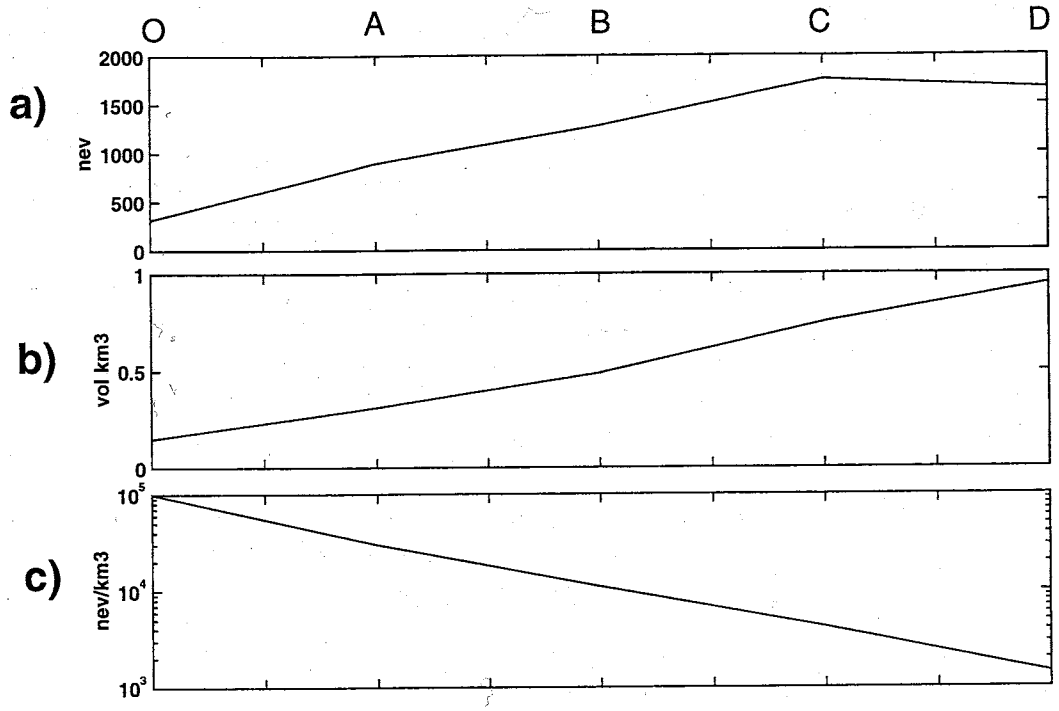
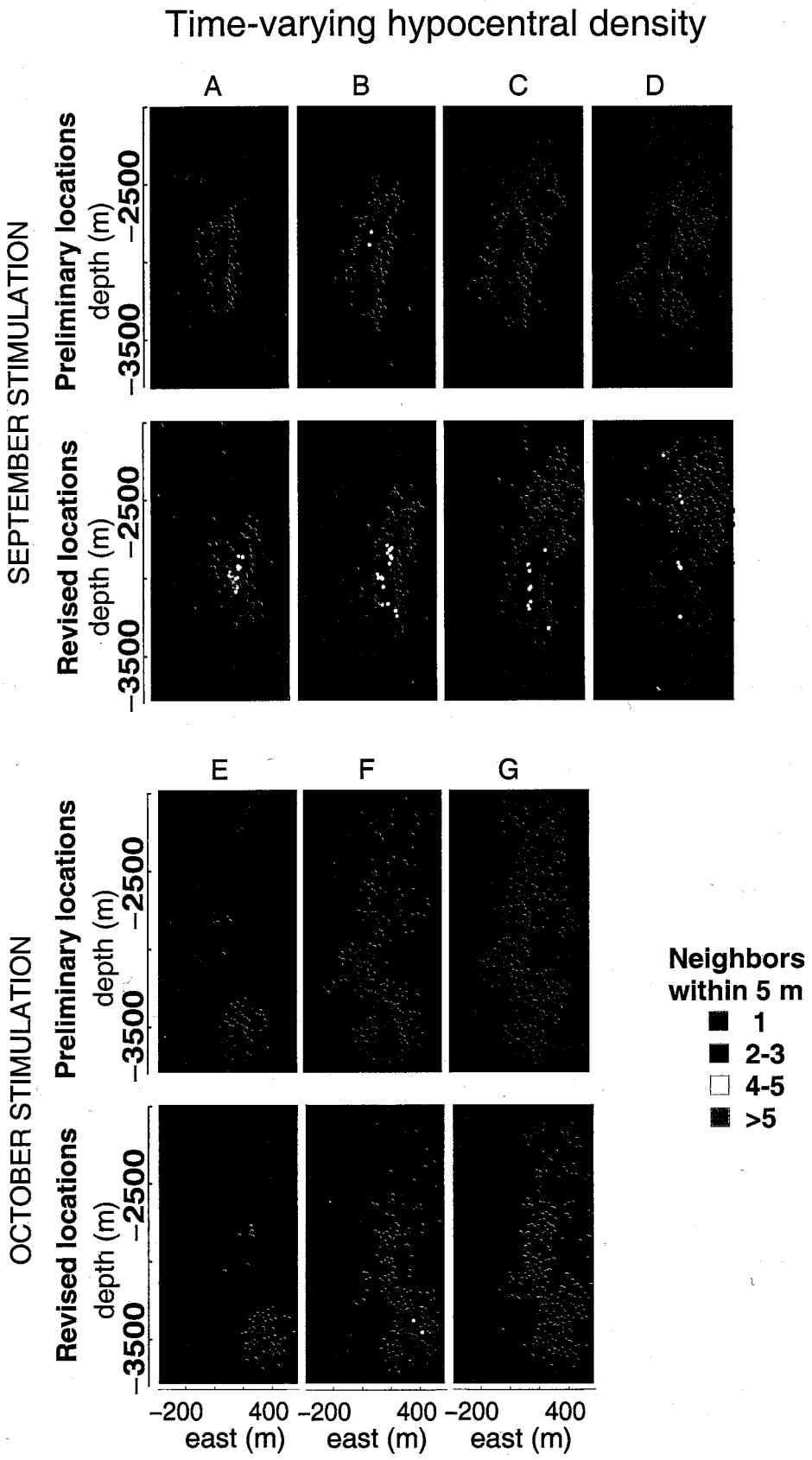


Figure 14: a) Event counts from the first five 72-hour time intervals (A-D) for the September 1993 stimulation, as indicated in Figure 12a, beginning at point **o** (1623 hours, 2 Sept.). Vertical axis represents the number of earthquakes, horizontal axis indicates time interval (Slight decrease in seismicity for the final data point is a reflection of the shorter time interval of active stimulation; reservoir deflation began during this interval.) b) Reservoir volume in each time segment. Vertical axis represents volume in km^3. The volume was approximated by estimating an envelope about the seismicity in the X, Y and Z directions as shown in Figures 14b-d, and calculating the mean volume for each time interval. c) Event density estimate for each time interval. Vertical axis indicates estimated number of events per km^3, plotted on a logarithmic scale

fracture planes along which earthquakes may have occurred. Roff et al. (1996) further investigated linear and planar features through analysis of waveform amplitude ratios, and Phillips et al. (1997) investigated cluster structures via manual hypocenter relocations and slant stack amplitudes. Starzec et al. (2000) found a correlation between the shapes of variograms of shear displacement and fracture orientation, suggesting a correlation of spatial variability of seismic slip and fracture complexity.

Tracking the highest concentrations of seismicity as a function of time may lead to the temporal mapping of flow paths - in essence, watching a movie of the reservoir development. We illustrate in Figure 15 an east-west cross-section of the hypocenters in each of the seven 72-hour intervals A-G identified in Figure 12a. Upper panels depict original locations; lower panels show relocated hypocenters. Because we required the existence of seven phase arrivals to relocate events, and the period prior to time segment (a) (beginning at point O, indicated on Figure 12a) offered no hydrophone arrivals, we are unable to compare relocations for that period. We have calculated hypocentral density by counting neighbors within a sphere of 5 m radius centered on each hypocenter during each of the time intervals. Using the method of Snoke (1987) to estimate source radius from corner frequencies, we have found that the median source radius for the relocated data subset to be approximately 3 m, so most events which fall within the 5 m search radius will have abutting or overlapping rupture patches. Upper panels of Figure 15 show results for the seven time intervals (A-G) using preliminary locations; lower panels illustrate relocated events. Hypocenters are color-coded to indicate spatial density, as indicated on the color key in Figure 15. Total numbers of earthquakes depicted are 6871 and 6217 for preliminary and relocated events, respectively; the discrepancy arises out of the number of

Figure 15: Spatially varying hypocentral density as a function of time. Preliminary (top panels) and revised (bottom panels) depicting seismic development of the Soultz reservoir during each of the time intervals A-D, and E-G of Figure 14a. Color represents the number of neighbors within a 5 m radius. Intervals A-D correspond to the September stimulation, when the open hole extended from 2800 to 3400 m depth and the bottom of the well was sanded. Intervals E-G correspond to October injection following removal of the sand.


Time-varying hypocentral density
A
B
C
D
SEPTEMBER STIMULATION
Preliminary locations
depth (m)
−2500
−3500
Revised locations
depth (m)
−2500
−3500
E
F
G
OCTOBER STIMULATION
Preliminary locations
depth (m)
−2500
−3500
Revised locations
depth (m)
−2500
−3500
−200
400
east (m)
−200
400
east (m)
−200
400
east (m)
Neighbors within 5 m
1
2-3
4-5
>5

events which were orphaned during the clustering process; these earthquakes were not repicked and so are not included in the relocations.

Overall, a clear increase in clustering after relocation is indicated in Figure 15. The tightest clustering remains nearest the injection point during the September stimulation (panels A-D), although the appearance of doublet events (magenta) on the edges of the developing reservoir margins suggests a concentration of activity at the pressure front of the migrating fluid. A similar, albeit smaller, trend is apparent in panels (E) and (F), which correspond to the later, October, stimulation following removal of sand from the bottom of the borehole. Figure 16 shows histograms of hypocenter density values to demonstrate the increased clustering following automatic repicking. The vertical axis of the histogram (event counts per cluster size) has been truncated so that smaller values may be displayed; single-event density counts are 5294 and 3755 for preliminary and relocated hypocenters, respectively. Among events in the seven time periods (A-G) displayed in Figure 15, 74% of relocated events do not locate within 5 m of their nearest neighbor, whereas 95% of original hypocenters fail to cluster. Table 4 presents the singlet (events which have no neighbors within a 5 m radius) versus multiplet (two or more events within a 5 m radius) statistics by time interval, comparing percentage of singlet events for preliminary and relocated hypocenters during the seven time periods of Figure 15.

Table 4: Percentages of singlet/multiplet events in 72-hour time segments

Time segment	A	B	C	D	E	F	G
% singlet preliminary	88	93	97	98	100	99	100
% multiplet preliminary	12	7	3	2	0	1	0
% singlet relocated	67	69	76	80	81	84	92
% multiplet relocated	33	31	24	20	19	16	8

We note that relocating events using adjusted phase picks may still not provide optimal relative and absolute hypocenter relationships, without addressing inadequacies of the velocity model and station corrections. The Soultz velocity model and station corrections were derived from a single downhole shot (e.g. Jupe et al., 1994) in combination with analysis of microseismicity using the preliminary phase picks. We observe systematic biases in these phase picks, which, if uncorrected, will contribute to the inaccuracy of preliminary velocity assumptions and subsequent analyses using those velocities. For example, pick times shown in Figure 5b demonstrate that the mean S-wave pick for station 4601 is clearly late, occurring more than a quarter cycle (~6 ms) after the S-wave arrival. Because we impose a zero mean constraint on our pick adjustments, this phase pick bias is an artifact of the preliminary pick distribution and has contributed to velocity and station correction estimates at Soultz. The bias may be accommodated in part by imposing *post priori* station correction, but this does not address the problem of separating actual pick bias from possible velocity heterogeneities. Removing this artifact from the catalogue using the waveform stacks should significantly improve efforts to refine velocity modelling at Soultz.

SUMMARY

We have demonstrated an automatic phase-repicking algorithm as applied to microearthquakes at the Soultz-sous-Forêts reservoir in the Rhine Graben of eastern France. The algorithm was tested using the results of careful, manual repicking on two isolated clusters of seismicity within the Soultz catalogue, then applied to the larger dataset.

Results of the relocation on the test subset demonstrate that the algorithm can recover detailed structures within the microseismicity cloud with a degree of improvement approaching that possible with careful, manual repicking. Median hypocenter discrepancies compared to Phillips' (2000) careful manual relocations improved from 31 m to 7 m upon adjusting the preliminary phase picks with our automated algorithm. In the cases of distorted or clipped waveforms, such as with the Soultz hydrophone, human intervention may still be necessary if optimal adjustment of every phase for every earthquake is desired.

Application to a data set of ~7000 Soultz microearthquakes results in significant enhancement of seismogenic fractures and faults within the reservoir, consistent with detailed studies on smaller selected subsets. Event associations with nearest neighbors, estimated by counting the number of earthquakes within a 5 m radius, increases from 5% to 26% of the catalogue. These linear joint and fracture structures tend for the most part to be steeply dipping, with predominant northwest-southeast orientations, although northeasterly-striking fracture sets also comprise a significant portion of the resolved structure. The previously diffuse cloud also exhibits subparallel, en echelon linear features with a more northerly strike, among shallow (2500 m - 2900 m) events. Between 2900 m and 3100 m we observe features which may represent intersecting faults or fractures, some dipping steeply to the east, others dipping steeply to the west. Overall, the image presented by relocated seismicity indicates that slip is occurring on fractures with varying orientations, whose characteristics change with depth. Evaluation of the relationships among these complex structures may yield better insights into reservoir flow behavior.

Correction of the phase arrival estimates to reduce random errors in the hypocenter location process can sharpen the results of subsequent applications including precise event location via SEL, JHD and/or collapsing techniques, as well as seismic tomography, reciprocal array processing, spatial b-value or fractal dimension analysis and other imaging techniques which depend upon highly consistent phase picks. Additionally, the correlation, intra-cluster adjustment, stacking and subsequent stack correlation illuminates systematic biases in the preliminary phase picks which have contributed to inaccuracies in the initial velocity models. Identification and removal of these systematic data errors will allow for better modelling of the reservoir.

ACKNOWLEDGEMENTS

The authors wish to thank D. Alde for help with data selection criteria. Preliminary phase picking for the full catalogue was undertaken by T. Thomas. H. Tobin and J. Schlue provided thorough and insightful reviews of the manuscript. This work was supported under the MTC (More Than Cloud) project of Niitsuma Laboratories, Tohoku University, Sendai, Japan, the National Science Foundation Office of Polar Programs grant #9419267 and Sandia National Laboratories.

REFERENCES

Aster, R.C. and J. Scott (1993), *Comprehensive characterization of waveform similarity in microearthquake data sets*, Bull. Seism. Soc. Am. *83*, 1307-1314.

Aster, R.C. and C.A. Rowe (2000), Automatic phase pick refinement and similar event association in large seismic datasets, in *Advances in Seismic Event Location*, C. Thurber and N. Rabinowitz (Editors), Kluwer, Amsterdam, 231-263.

Baria, R., J. Baumgartner, A. Gerard, R. Jung and J. Garnish (1999), *European HDR research programme at Soultz-sous-Forets (France) 1987-1996*, Geothermics *28*, 655-669.

Block, L. (1991). *Joint Hypocenter-Velocity Inversion of Local Earthquake Arrival Time Data in Two Geothermal Regions*, Ph.D. Dissertation, Massachusetts Institute of Technology, 448 pp.

Block, L.V., C.H. Cheng, M.C. Fehler and W.S. Phillips (1994), *Seismic imaging using microearthquakes induced by hydraulic fracturing*, Geophysics *59*, 102-112.

Cornet, F.H. and R. Jones (1994), *Field evidence on the orientation of forced water flow with respect to the regional principal stress directions*, Proc. 1st North Amer. Rock Mech. Symp., Austin, Texas, 61-69.

Deichmann, N. and M. Garcia-Fernandez (1992), *Rupture geometry from high-precision relative hypocentre locations of microearthquake clusters*, Geophys. J. Int. *110*, 501-517.

Dodge, D.A., G.C. Beroza and W.L. Ellsworth (1995), *Foreshock sequence of the 1992 Landers, California earthquake and its implications for earthquake nucleation*, J. Geophys. Res. *100*, 9865-9880.

Fehler, M., L. House and H. Kaieda (1987), *Determining planes along which earthquakes occur: Method and application to earthquakes accompanying hydraulic fracturing*, J. Geophys. Res. *92*, 9407-9414.

Fehler, M., W.S. Phillips, R. Jones, L. House, R. Aster and C. Rowe (2000), *A method for improving relative earthquake locations*, Bull. Seism. Soc. Am. *90*, 775-780.

Fremont, M.-J. and S. D. Malone (1987), *High precision relative locations of earthquakes at Mount St. Helens, Washington*, J. Geophys. Res. *92*, 10,223-10,236.

Genter, A. and H. Traineau (1996), *Analysis of macroscopic fractures in granite in the HDR geothermal well EPS-1, Soultz-sous-Forets, France*, *J.* Volc. Geotherm. Res. *72*, 121-141.

Gillard, D., A.M. Rubin and P. Okubo (1996), *Highly concentrated seismicity caused by deformation of Kilauea's deep magma system,* Nature *384,* 343-346.

Goldstein, P., D. Dodge, M. Firpo and S. Ruppert (1998), *What's new in SAC2000? Enhanced processing and database access,* Seism. Res. Lett. *69*, 202-205.

Got, J.-L., J. Frechet and F.W. Klein (1994), *Deep fault plane geometry inferred from multiplet relative relocation beneath the south flank of Kilauea,* J. Geophys. Res. *99,* 15,375-15,386.

Harris, M. and C. Young (1997), *MatSeis: a seismic GUI and tool-box for MATLAB,* Seism. Res. Lett. *68*, 267-269.

Ito, A. (1995), *High resolution relative hypocenters of similar earthquakes by cross-spectral analysis method,* J. Phys. Earth *33*, 279-294.

Jones, R.H. and R.C. Stewart (1997), *A method for determining significant structures in a cloud of earthquakes*, J. Geophys. Res. *102,* 8245-8254.

Jones, R.H., A. Beauce, A. Jupe, H. Fabriol and B.C. Dyer (1995), *Imaging induced microseismicity during the 1993 injection tests at Soultz-sous-Forets, France,* Proc. World Geotherm. Cong., Florence, 2665-2669.

Jung (1991), *Hydraulic fracturing and hydraulic testing in the granitic section of Borehole GPK1, Soultz-sous-Forets*, Geotherm. Sci. Technology. *3*, 149-198.

Jupe, A.J., R.H. Jones, J. Willis-Richards, B. Dyer, J. Nicholls and P. Jacques (1994), *Report on HDR Phase 4 - Soultz Experimental Programme 1993/1994, CSM Associates Ltd.*, IR02/12.

Jupe, A.J., R. Jones, B. Dyer and S. Wilson (1998), Monitoring and management of fractured reservoirs using induced microearthquake activity, Soc. Petroleum. Engineers contribution SPE 47315, in *International Collaborative Project on Mapping of Engineered Geothermal Reservoirs*, MTC Project Report,.

Kissling, E. (1988), *Geotomography with local earthquake data*, Rev. Geophys. *26*, 659-698.

Lance, G.N. and W.T. Williams (1967), *A general theory for classificatory sorting strategies 1. Hierarchical systems.*, Computer J. *10*, 271-276.

Lees, J.M. (1998), *Multiplet analysis at Coso geothermal,* Bull. Seism. Soc. Am. *88*, 1127-1143.

Ludwig, J.A. and J.F. Reynolds (1988), *Statistical Ecology, a Primer on Methods and Computing,* John Wiley & Sons, New York, 337 pp.

Moriya, H., K. Nakazato and H. Niitsuma (2000), *Detailed fracture system of Soultz-sous-Forets HDR field evaluated by using microseismic doublet/multiplet,* Pure Appl. Geophys., submitted.

Nadeau, R., W. Foxall and T. McEvilly (1995), *Clustering and periodic recurrence of microearthquakes on the San Andreas Fault at Parkfield, California*, Science *267*, 503-507.

Nanjo, K., H. Hagahama and M. Satomura (1998), *Rates of aftershock decay and the fractal structure of active fault systems*, Tectonophysics *287*, 173-186.

Park, J., C.R. Lindberg and F.L. Vernon III (1987), *Multitaper spectral analysis of high-frequency seismograms,* J. Geophys. Res. *92*, 12,675-12,684.

Parker, R.L. and M.K. McNutt (1980), *Statistics for the one-norm misfit measure*, J. Geophys. Res. *85*, 4429-4430.

Phillips, W.S., L.S. House and M.C. Fehler (1997), *Detailed joint structure in a geothermal reservoir from studies of induced microearthquake clusters,* J. Geophys. Res. *102*, 11,745-11,763.

Phillips, W.S. (2000), *Precise microearthquake locations and fluid flow in the geothermal reservoir at Soultz-sous-Forets, France,* Bull. Seism. Soc. Am. *90*, 212-228.

Phillips, W.S., J.T. Rutledge, L.S. House and M.C. Fehler (2000), *Induced microearthquake patterns in hydrocarbon and geothermal reservoirs,* Pure Appl. Geophys., submitted.

Poupinet, G., W.L. Ellsworth and J. Frechet (1984), *Monitoring velocity variations in the crust using earthquake doublets: An application to the Calaveras Fault, California,* J. Geophys. Res. *89,* 5719-5731.

Pujol, J. (1992), *Joint hypocentral location in media with lateral velocity variations and interpretation of the station corrections,* Phys. Earth Planet. Interiors *75*, 7-24.

Roff, A., W.S. Phillips and D.W. Brown (1996), *Joint structures determined by clustering microearthquakes using waveform amplitude ratios, I*nt. J. Rock Mech. Min. Sci. Geomech. Abstr. *33,* 627-639.

Rowe, C.A., R.C. Aster, M.C. Fehler, W.S. Phillips and P.R. Kyle (1998), *Hypocenter Relocation in Tectonic and Volcanic Seismic Data Sets Using Automatic phase Cor-*

relation: Application of Cross-correlation, Signal Envelope and Principal Components Methods to Improve Pick Consistency, EOS *79*, F588.

Rowe, C.A. and R.C. Aster (1999). *Application of automatic, adaptive filtering and eigenspectral techniques to large digital waveform catalogs for improved phase pick consistency and uncertainty estimates*, EOS *80*, F660.

Rowe, C.A., R. C. Aster, B. Borchers and C. J. Young (2000), *Relocation of seismicity associated with the 1989-1990 eruption of Redoubt Volcano, Alaska. Part I: An automatic, adaptive algorithm for refining phase picks in large seismic data sets*, J. Geophys. Res., in preparation.

Rubin, A. M., D. Gillard and J.-L. Got (1998), *A reinterpretation of seismicity associated with the January 1983 dike intrusion at Kilauea Volcano, Hawaii*, J. Geophys. Res. *103*, 10,003-10,015.

Rubin, A.M., D. Gillard and J.-L. Got (1999), *Streaks of microearthquakes along creeping faults,* Nature *400*, 635-641.

Shearer, P.M. (1997), *Improving local earthquake locations using the L1 norm and waveform cross-correlation: Application to the Whittier Narrows, California, aftershock sequence,* J. Geophys. Res. *102*, 8269-8283.

Shearer, P.M. (1998), *Evidence from a cluster of small earthquakes for a fault at 18 kim depth beneath Oak Ridge, Southern California*, Bull. Seism. Soc. Am. *88*, 1327-1336.

Snoke J A (1987), *A stable determination of (Brune) stress drops*, Bull. Seism. Soc. Am. *77*, 530-538.

Spudich, P. and T. Bostwick (1987), *Studies of the seismic coda using an earthquake clus-*

ter as a deeply buried seismograph array, J. Geophys. Res. *92*, 10,526-10,546.

Thomson, J.D. (1982), *Spectrum estimation and harmonic analysis,* Proc. IEEE, *70,* 1055-1096.

Thurber, C., S. Roecker, T. McEvilly, F. Haslinger, D. Steinberg, and W. Ellsworth (2000), *Earthscope at Parkfield, California: setting the stage for drilling,* EOS *81*, S318,.

Waldhauser, F. W.L. Ellsworth and A. Cole (1999), *Slip-parallel seismic lineations on the northern Hayward Fault, California*, Geophys. Res. Lett. *26,* 3525-3528.

Wiemer, S. and M. Wyss (1997), *Mapping the frequency-magnitude distribution in asperities: An improved technique to calculate recurrence times?,* J. Geophys. Res. *102,* 15,115-15,128.

APPENDIX 2: DATA USED IN THIS STUDY

Waveform and preliminary phase picks for the two applications in Parts B and C may be obtained from the following sources:

Redoubt Volcano, Alaska

Event-detected waveform data for the 1989-1990 Redoubt eruption is part of the Alaska Volcano Observatory (AVO) archive. These data exist on CD-ROM at the USGS AVO offices in Anchorage, Alaska. Data requests may be directed to Dr. John Power (jpower@usgs.gov). The waveforms are in AH format. Preliminary phase data and preliminary hypocenters are stored in HYPOELLIPSE (Lahr, 1989[1]) format and are available separately. HYPOELLIPSE control files used by AVO for event location at Cook Inlet volcanoes may likewise be requested from AVO. Pickfiles and waveforms for 1989-1994 occupy seven CD-ROMs, which also include data from Cook Inlet volcanoes Spurr, Iliamna and Augustine.

Soultz geothermal reservoir, France

Waveforms and picks for the Soultz reservoir data may not be generally distributed except with the express agreement of the MTC (More Than Cloud) program which is directed by Niitsuma Laboratories, Tohoku University, Sendai, Japan. These data may be requested by contacting Professor Hiroaki Niitsuma, ni@ni4.earth.tohoku.ac.jp

[1] Lahr, J. C., *HYPOELLIPSE / Version2.0: A computer program for determining local earthquake hypocentral parameters, magnitude, and first motion pattern*, USGS. Open File Report 89-116, 92 pp., 1989.