Bulletin of the Seismological Society of America, Vol. 95, No. 6, pp. 2446–2461, December 2005, doi: 10.1785/0120040221

Waveform Cross-Correlation-Based Differential Travel-Time Measurements at the Northern California Seismic Network

by David P. Schaff and Felix Waldhauser

Abstract We processed the complete digital seismogram database for northern California to measure accurate differential travel times for correlated earthquakes observed at common stations. Correlated earthquakes are earthquakes that occur within a few kilometers of one another and have similar focal mechanisms, thus generating similar waveforms, allowing measurements to be made via cross-correlation analysis. The waveform database was obtained from the Northern California Earthquake Data Center and includes about 15 million seismograms from 225,000 local earthquakes between 1984 and 2003. A total of 26 billion cross-correlation measurements were performed on a 32-node (64 processor) Linux cluster, using improved analysis tools. All event pairs with separation distances of 5 km or less were processed at all stations that recorded the pair. We computed a total of about 1.7 billion *P*-wave differential times from pairs of waveforms that had cross-correlation coefficients (CC) of 0.6 or larger. The *P*-wave differential times are often on the order of a factor of ten to a hundred times more accurate than those obtained from routinely picked phase onsets. 1.2 billion *S*-wave differential times were measured with CC $\geq$ 0.6, a phase not routinely picked at the Northern California Seismic Network because of the noise level of remaining *P* coda. We found that approximately 95% of the seismicity includes events that have cross-correlation coefficients of CC $\geq$ 0.7 with at least one other event recorded at four or more stations. At some stations more than 40% of the recorded events are similar at the CC $\geq$ 0.9 level, indicating the potential existence of large numbers of repeating earthquakes. Large numbers of correlated events occur in different tectonic regions, including the San Andreas Fault, Long Valley caldera, Geysers geothermal field and Mendocino triple junction. Future research using these data may substantially improve earthquake locations and add insight into the velocity structure in the crust.

Introduction

One of the most fundamental datasets in seismology is the set of measured arrival times of various phases on a seismogram. These basic data are used to solve for earthquake hypocenters and also to derive velocity models or empirical travel-time curves. But there is an error associated with each measurement. *P*-wave arrival times at the Northern California Seismic Network (NCSN) are picked with an average pick error on the order of 0.1 sec. These errors map into significant scatter in the earthquake locations and reduce the resolution of tomographic images of the velocity structure.

It has long been established that cross-correlation measurements of differential travel times can improve these errors by an order of magnitude or more if the waveforms are similar. The differential travel times can be inverted directly for earthquake locations (e.g., Waldhauser and Ellsworth, 2000) or they can be inverted at each station to improve the absolute arrival times (e.g., Shearer *et al.*, 2005). Similar waveforms are produced when earthquakes have the same rupture mechanism and are colocated (i.e., share the same ray paths between source and receiver). Figure 1 shows an example of 38 virtually identical waveforms of a repeating earthquake sequence recorded at the NCSN station JST. In this case the event hypocenters are colocated within the location errors and the estimated source areas overlap such that it is assumed that the same asperity is rerupturing each time. For such similar waveforms, relative phase arrival times can be obtained with subsample precision (Poupinet *et al.*, 1984). With a sampling rate of 100 samples/sec, errors in relative arrival time measurements are less than 1 msec in the optimal case. Cross-correlation measurements are particularly important for *S* waves, because *S*-wave onsets are often obscured by the *P*-wave coda and as a result are rarely picked for NCSN data. *S*-wave measurements are especially

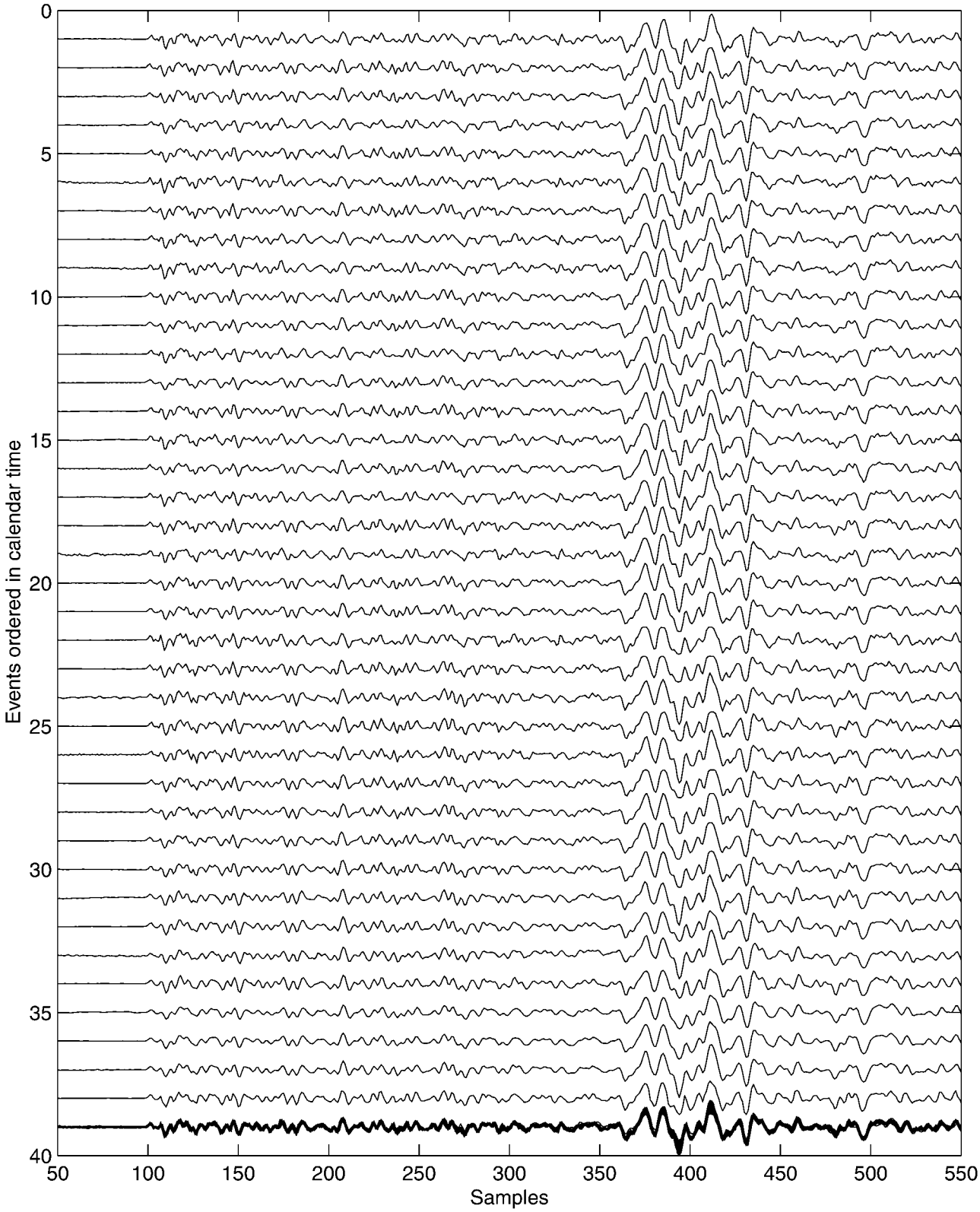

Figure 1. Waveforms of 38 repeating earthquakes on the Calaveras Fault recorded at station JST. Bottom trace shows all 38 waveform superposed. Note the high similarity between all waveforms. From Schaff and Beroza (2004).

important to better constrain depths in earthquake locations. Even when earthquakes are not exactly colocated, waveforms can be similar enough to provide significant improvement in relative arrival-time measurements over ordinary phase picks, as long as the focal mechanisms are similar as well as the frequency content of the *P* and *S* phases.

The use of waveform cross correlation has a long history of identifying similar events and improving earthquake location for small case examples (Poupinet *et al.*, 1984; Fréchet, 1985; Ito, 1985; Frémont and Malone, 1987; Deichmann and Garcia-Fernandez, 1992; Got *et al.*, 1994; Dodge *et al.*, 1995; Nadeau *et al.*, 1995; Shearer, 1997; Lees, 1998; Phillips, 2000; Moriya *et al.*, 2003). After the realization that correlation data could be applied effectively for longer event separation distances, it has been combined with improved relative location techniques to relocate large numbers of events at spatial scales of more than a few kilometers (e.g., Rubin *et al.*, 1999; Waldhauser *et al.*, 1999; Waldhauser and Ellsworth, 2000; Rowe *et al.*, 2002; Schaff *et al.*, 2002; Waldhauser *et al.*, 2004; Shearer *et al.*, 2005; Hauksson and Shearer, 2005). Such larger scale applications were greatly helped by the increasing amount of high-quality digital waveform data in areas of dense seismicity, and the substantial increase in computational power and storage capability to handle and process that data. In this article we present procedures and results from measuring differential times via waveform cross correlation on a massive scale at the NCSN, a seismographic network of ~900 stations that records ~10,000 events each year. The uniformity and consistency of the measurements make these data not only useful for location purposes, but they also present a comprehensive characterization of waveform similarity across various tectonic regions in northern California.

Data and Methods

The complete digital seismogram database of the Northern California Earthquake Data Center (NCEDC), totaling 225 GB of compressed waveform data (~700 GB in uncompressed form), has been made available to us on 10 DLT tapes (D. Neuhauser, personal comm., 2001; Neuhauser *et al.*, 1994). The seismograms are from about 225,000 events in northern California, recorded at a total of 900 short-period, vertical-component stations of the NCSN, during the period of March 1984 (when digital recording began) through May 2003. The original data was stored in a compressed binary CUSP format (http://quake.geo.berkeley.edu) that was then converted to the SAC format (http://www.l1nl.gov/sac) for processing. Each header was updated with theoretical *P*- and *S*-wave travel-time information when phase picks were not available. The 15 million SAC seismograms were then reorganized from a calendar ordering to a station ordering, as correlation measurements are performed on a station-by-station basis. Disk operations and network transfer rates were on the order of 1 MB/sec. Therefore each disk access operation amounted to about 3 days if uninterrupted—copying the data from the DLT drive, uncompressing, converting from CUSP to SAC, recompressing, and reorganizing into station subdirectories. The DLT tapes had to be manually changed after each extraction. The total amount of time involved for data handling and manipulation and development of associated software was about 2 months.

Event Pair Selection Based on Double-Difference Locations

Since waveform similarity breaks down with increasing interevent separation distance, we implemented an event separation threshold to select event pairs suitable for cross correlation. To improve the accuracy of interevent distances from which we determine such pairs of events we have relocated about 240,000 events by applying the double-difference algorithm hypoDD (Waldhauser and Ellsworth, 2000; Waldhauser, 2001) to a total of about 5 million NCSN *P*-phase picks. Using these improved locations, we chose an interevent distance threshold of 5 km. This selection is based in part on the quarter wavelength rule (Geller and Mueller, 1980), which describes the rapid decrease of waveform similarity with increasing interevent distance. The threshhold also accounts for remaining errors in the double-difference locations (~100 m) and the larger errors (~1 km) in events not relocated by the double-difference method due to lack of good station coverage. As shown later, the 5-km threshold is a good compromise between catching most of the correlated events and keeping the computational time at a reasonable level.

Massive-Scale Cross Correlation

Differential travel times were computed using a modified version of the cross-correlation algorithm described in Schaff *et al.* (2004). The modifications include the use of a correlation detector rather than a correlation function in order to recover time lags greater than half the window length. When dealing with finite duration signals, time-domain cross correlation is computed by fixing one window on the first seismogram and moving a sliding window over the second seismogram padded with zeros (Fig. 2). In functional form, for a discrete time series of length N, the cross-correlation function for lags, τ, is:

$$C(\tau) = c \sum_{n=0}^{N-1} y_1(n) y_2(n + \tau),$$

where $1 - N \le \tau \le N - 1$,

$$c = \left[\sum_{n=0}^{N-1} y_1^2(n) \sum_{n=0}^{N-1} y_2^2(n) \right]^{-1/2},$$

$$\text{and} \quad y_2(i) = 0 \begin{cases} i < 0 \\ i > N - 1 \end{cases}.$$

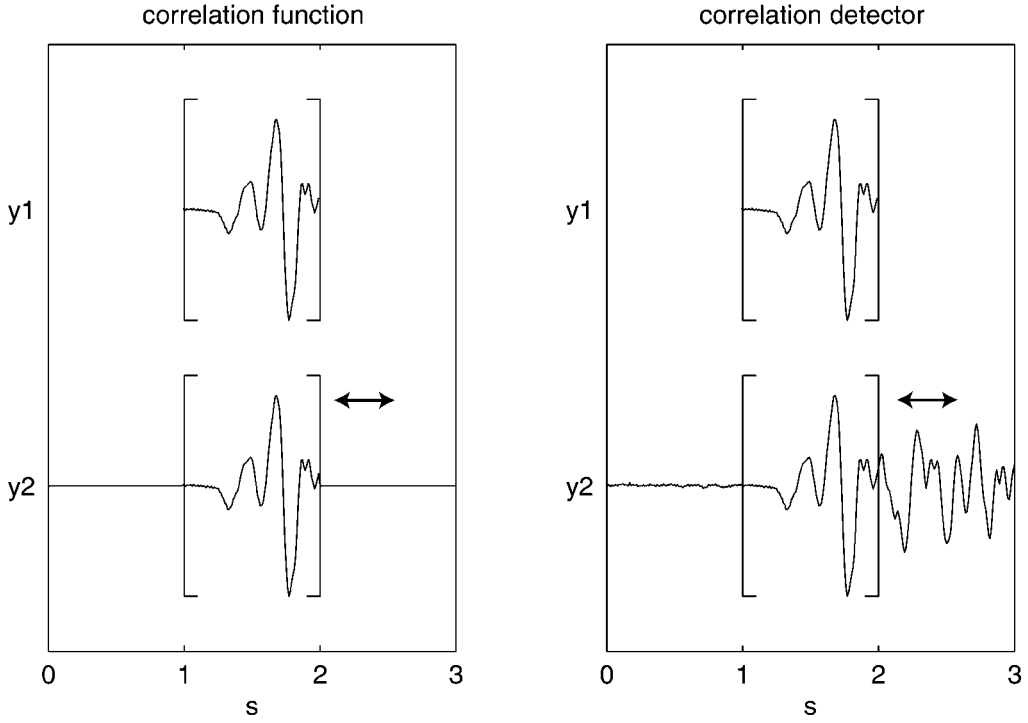

Figure 2. Illustration of the difference between a correlation function and a correlation detector. Window (shown by brackets) is fixed for seismogram y1. Sliding window for seismogram y2 runs over zero padding for the correlation function and over data for the detector.

An equivalent result is obtained if the cross-correlation function is first computed in the frequency domain as the cross spectrum and then inverse Fourier transformed back into the time domain. Although the cross-correlation function is technically defined for lags plus and minus the window length, in practice only lags less than or equal to half the window length can be recovered (Schaff *et al.*, 2004). The reason is beyond this point, the percentage of similar energy in the two windows is less than 50% due to less than half the window lengths overlapping. A related effect is that the cross-correlation coefficient (CC) measurement degrades with increasing initial offset of the two seismograms (Schaff *et al.*, 2004). If instead of padding with zeros, the original data is retained in the second seismogram, both of these negative effects with correlation functions are eliminated (Fig. 2). We call such an application employing a correlation detector. Now the sliding window can align arbitrarily long offsets and perfectly capture the correct correlation coefficient. Note: this approach is basically equivalent in procedure to commonly used correlation-based event detectors that try to pick out similar events from continuous data streams. The difference in this case is we instead only window around the *P* or S phases of interest.

Many waveform cross-correlation methods employ correlation functions or cross-spectral techniques and are therefore limited by these fixed-window-length, finite-duration records in their ability to recover large offsets and faithfully measure the similarity. Cross-spectral techniques are even less capable of recovering large offsets (only a fraction of the window length compared to one half for correlation functions) because smoothing of the rapidly increasing phase in the complex plane biases the estimate toward zero (Schaff *et al.*, 2004). Note: some authors have employed a two-step procedure to partially overcome these limitations by first computing the correlation function, realigning the seismograms, and then doing a final delay estimate to subsample precision as well as obtaining a more representative similarity or error measurement (Rowe *et al.*, 2002; Schaff *et al.*, 2004). This is a practical issue since the initial windows may be offset by substantial amounts due to mispicks or theoretical travel times. For example, if two seismograms were mispicked each by 0.5 sec, the total offset could range up to 1 sec. A correlation function (obtained by zero-padding in the time domain or computed by the inverse Fourier transform of the cross spectrum) using 1-sec window lengths would not be able to align these offsets. Figure 3 shows examples of automatically determined *P*-wave arrival-time adjustments of similar events observed at station JST after application of a correlation detector that incorporates the real data before and after the initial window instead of padding with zeros. These *P*-wave trains have CC ≥ 0.9 and adjustments > 0.9 sec for window lengths of 1 sec.

A battery of tests was conducted to define correlation parameters that produce robust delay time measurements in

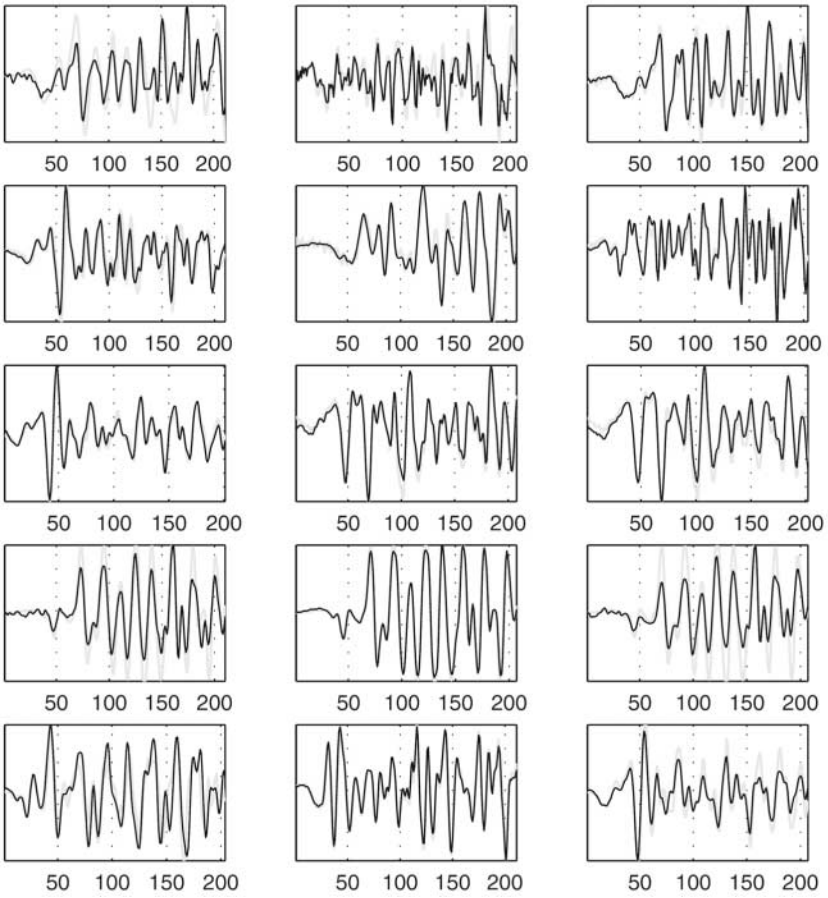

Figure 3. A few examples of aligned *P* waves for several pairs of events (black and gray overlaying seismograms) obtained from a correlation detector, which would have been missed by an ordinary correlation function. All adjustments are > 0.9 sec, which is more than half the window length of 1 sec. The *P*-wave trains are very similar with CC ≥ 0.9. *X*-axes are in samples (delta $= 0.01$ sec.).

an efficient manner. Based on these tests, of which results will be shown later, we chose the following input parameters to run uniformly across the entire network.

- Seismograms were filtered from 1.5 to 15 Hz (the instrument is reliable in this band).
- Correlation measurements were made for both 1- and 2-sec window lengths, for both *P*- and *S*-wave trains.
- The lags searched over were $\pm$ 1 sec.
- *P*-wave windows are initially aligned on phase-pick data if available from the network bulletin, or on theoretical travel times based on a 1D velocity model. *S*-wave travel times are computed as 1.732 times the *P*-wave travel time.

A total of 26 billion cross-correlation measurements (*P* and *S* waves, two windows each) were performed. The computations were performed on a 32 node Linux cluster, each node equipped with two 1.2 GHz Athlon processors, 1 GB of fast RAM, and 20 GB of scratch space. Cross correlations were performed at a rate of about 10 million measurements per CPU hour. (Note: for these window lengths and lags this is a factor of 10 faster than our earlier algorithm that computed the cross-correlation function first as the cross spectrum and then inverse Fourier transformed back into the time domain.) Total processing time including input/output operations was approximately two weeks. A RAID Tb storage system was used to store the waveforms and the measurement output. Since the correlations operate on a station by station basis, they are naturally parallelizeable and can use any number of free processors. Binary output files were saved for each station with the event pair indices and differential travel-time and correlation-coefficient measurements for both 1- and 2-sec window lengths. All data with CCs of 0.6 and above were saved, resulting in 1.7 billion *P*-wave and 1.2 billion *S*-wave correlation pair measurements for both window lengths. Note that *S* phases are not routinely picked by the NCSN. As shown later in this article and in previous work on waveform-based event relocation in northern California (Waldhauser and Ellsworth, 2000; Schaff *et al.*, 2002; Waldhauser *et al.*, 2004), the most robust measurements are made for *P*-wave trains that correlate above the CC $= 0.7$ threshold, but useful measurements are sometimes possible at the CC $= 0.6$ level or even lower, for example for repeating events where the underlying waveforms are similar but have random environmental noise superimposed. The resulting differential travel-time database is about 63 GB in size.

Results

An overview of the cross-correlation results is given in Figure 4. Shown are all events recorded at the NCSN between

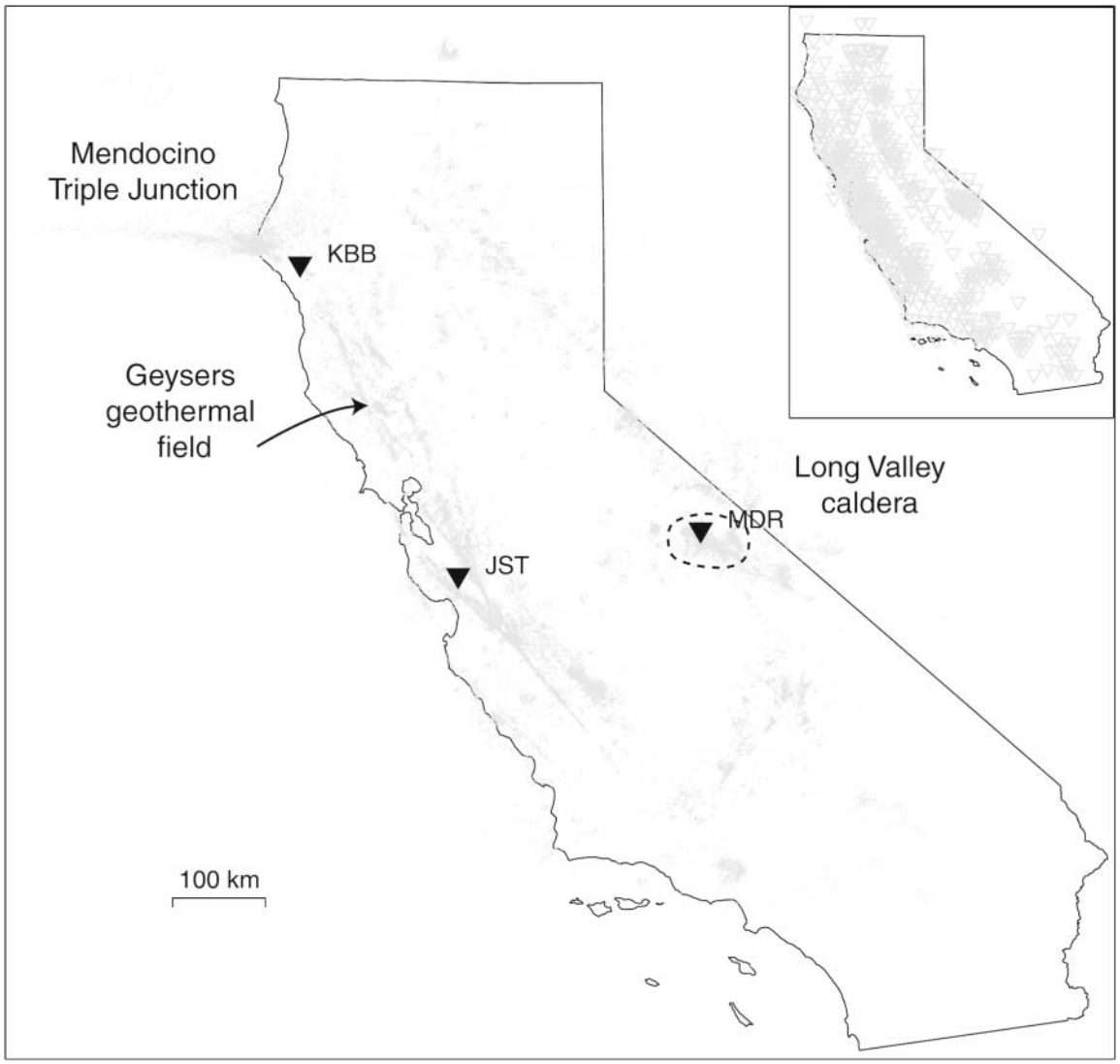

Figure 4. All events between 1984 and 2003 that have *P*-wave cross-correlation coefficients of CC $\geq$ 0.7 with at least one other event at four or more stations. These ~210,000 events represent 95% of the total seismicity with waveforms available. Event hypocenters are plotted at improved double-difference locations using phase data alone. Inset shows stations where correlation measurements were made.

1984 and 2003 that have similar *P*-wave trains at the CC $\geq$ 0.7 level (1-sec window lengths) with at least one other event at four or more stations. The ~200,000 events represent 95% of the total number of events for which waveforms are available. Ninety percent of the total number of events share similar *P*-wave trains with at least one other event at 8 or more stations, and 82% at 12 or more stations. To get some idea of the station coverage constraining these events we compute the azimuthal gap (GAP) and the difference in range of distances to the station (DRANGE = max(dist) $-$ min(dist)), excluding the ~13,000 Mendocino events because they are offshore and outside the network (GAP $>$ 180°). Eighty-six percent of the events have CC $\geq$ 0.7 at a minimum of four stations with GAP $\leq$ 180° and DRANGE $\geq$ 10 km. For CC $\geq$ 0.7 at a minimum of eight stations with GAP $\leq$ 180° and DRANGE $\geq$ 20 km, the percentage of events meeting these criteria is 83%. These surprisingly high numbers indicate that a large percentage of the NCSN catalog can be relocated by substituting ordinary phase picks with accurate differential travel times obtained from waveform cross correlation.

Areas with large numbers of highly correlated events can better be identified in Figure 5. This figure shows the percentage of events, within bins of 5 $\times$ 5 km, that have CC $\geq$ 0.7 for *P*-wave trains with at least one other event at four or more stations. A threshold of 0.7 is chosen from our experience in relocating events in California that balances the tradeoff of trying to increase the number of observations and events that correlate while also reducing the risk of introducing large outliers for these window lengths, with the

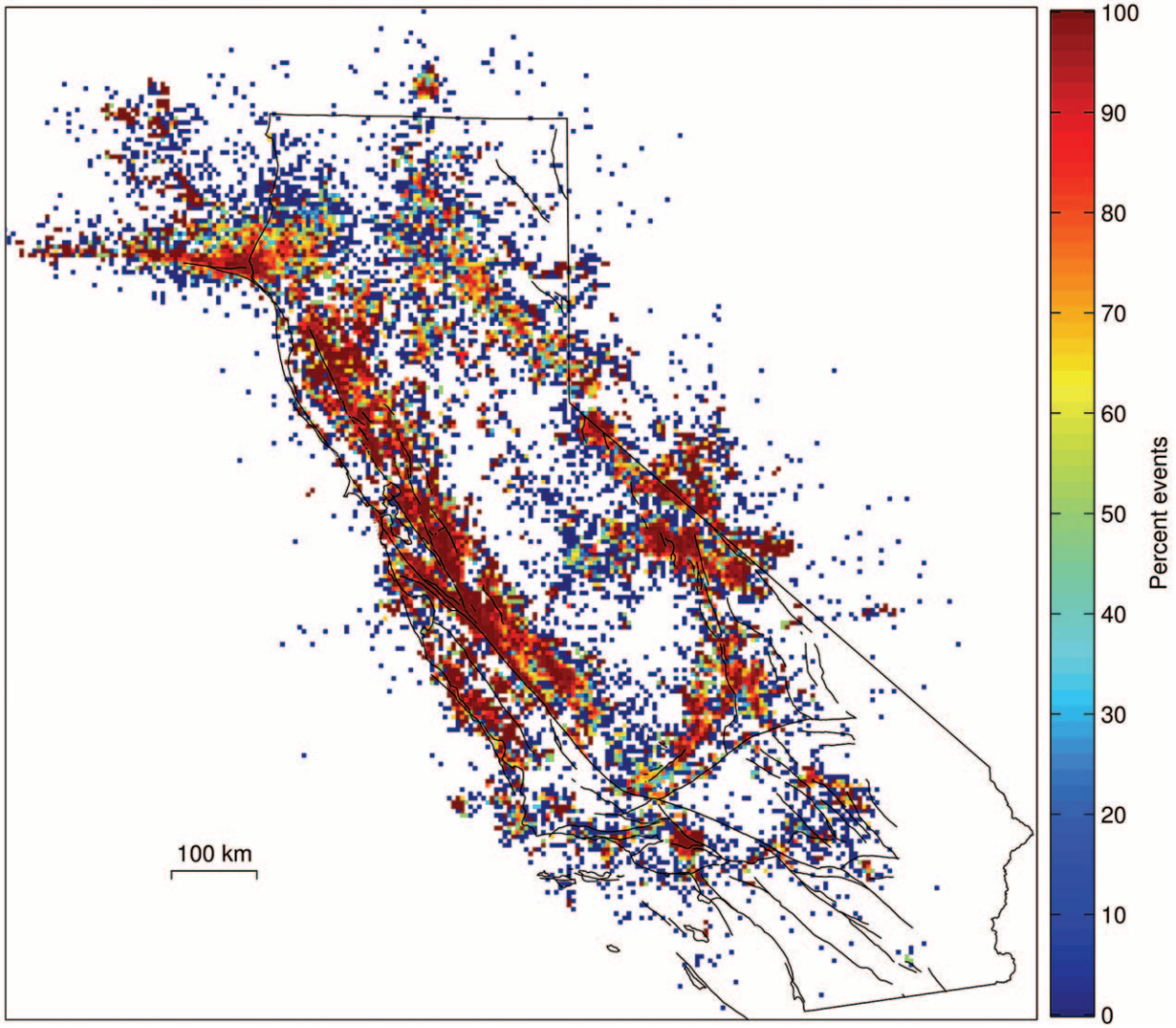

Figure 5. Percentage of correlated events that have cross-correlation coefficients of CC $\geq$ 0.7 with at least one other event recorded at four or more stations. Percentages are computed from the total number of events within bins of 5 $\times$ 5 km.

overall goal of substantially improving the locations for the largest percentage of events. (A similar figure is obtained if only *S*-wave trains are considered.) Figure 5 reveals that greater than a 75% level is obtained for much of the area in different tectonic settings such as the San Andreas Fault (SAF) system, the Long Valley caldera, the Geysers geothermal field, and the Mendocino triple junction. To first order, the areas of highly correlated events agree with areas of dense seismicity, suggesting that the closeness of events is the main factor for producing similar waveforms. Variations in focal mechanisms will also play a role, but it is difficult to separate the two effects.

A similar picture is obtained when the percentage of correlated events is plotted for individual stations that recorded them (Fig. 6a). Forty percent of the stations plotted have at least 60% of their events correlating at the CC $\geq$ 0.7 level with at least one other event. The stations that recorded many correlated events are located along the SAF, at Geysers geothermal field, and in the Long Valley area (Fig. 6a). Figure 6b indicates the percentage of the total number of cross-correlation measurements that have *P*-wave correlation coefficients of CC $\geq$ 0.7. Again, stations that recorded events from creeping faults along the SAF system (e.g., Calaveras Fault or the Parkfield section of the SAF) have significantly larger percentages of correlated waveforms than stations that record seismicity in areas that are seismically less active.

Three Example Stations: JST, MDR, and KBB

Figure 7 shows the detailed results for station JST. At this particular station, which recorded 35,000 events from the SAF system (Fig. 7a), 40% of the events have at least one other event with CC $>$ 0.9 (62% for CC $\geq$ 0.8, 77% for CC $\geq$ 0.7) (Fig. 7b). The percentages of similar events at high CC thresholds observed at station JST are surprisingly high, but they include known areas of repeating events on

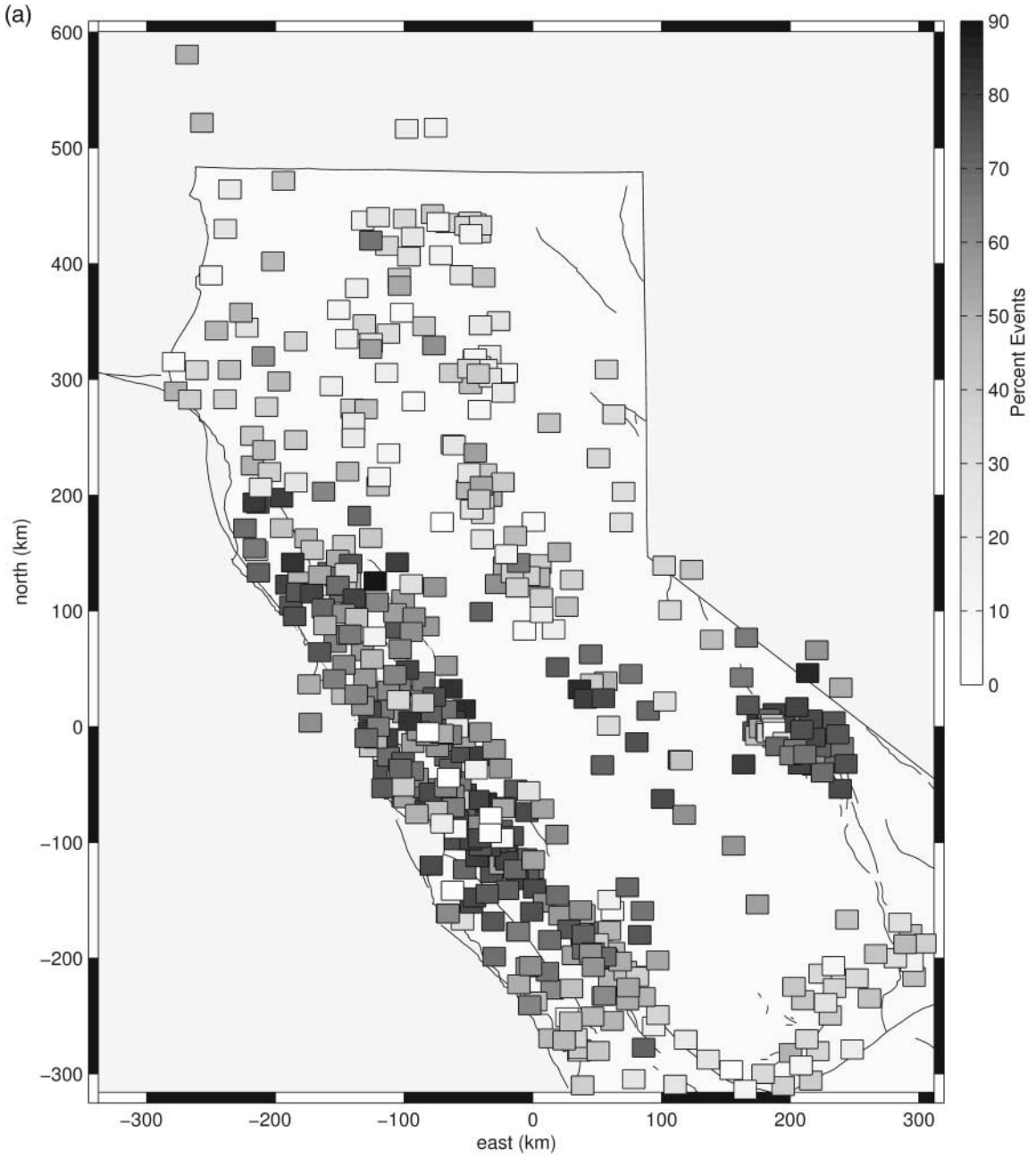

Figure 6. Station locations with percentages indicated for (a) number of correlated events recorded at each station, and for (b) number of cross-correlation measurements with CC $\geq$ 0.7 out of all measurements performed at that station. (*continued*)

the Calaveras and San Andreas faults. For a station in Long Valley Caldera (MDR) recording 72,000 events, the distribution is 18% for CC $\geq$ 0.9, 43% for CC $\geq$ 0.8, and 67% for CC $\geq$ 0.7. A station including 20,000 events in the different tectonic settings of Mendocino triple junction and Geysers geothermal fields yields correlation measurements where 19% of the events have at least one other event with CC $\geq$ 0.9, 36% with CC $\geq$ 0.8, and 57% with CC $\geq$ 0.7. The lower numbers of correlated events observed at the latter two stations most likely reflect the different, and probably more complex, faulting processes that take place in these areas, compared to the (mostly) strike-slip events recorded at JST. Table 1 and Table 2 summarize the number of measurements for the three stations and different thresholds. It is seen from Table 1 that a large percentage of the events correlate well across a variety of tectonic regions.

Not all the seismograms associated with an event have *P*-wave picks perhaps due to weak onsets or low signal-to-noise ratios. If we use theoretical initial *P*-wave window alignments for these events based on raytracing through a

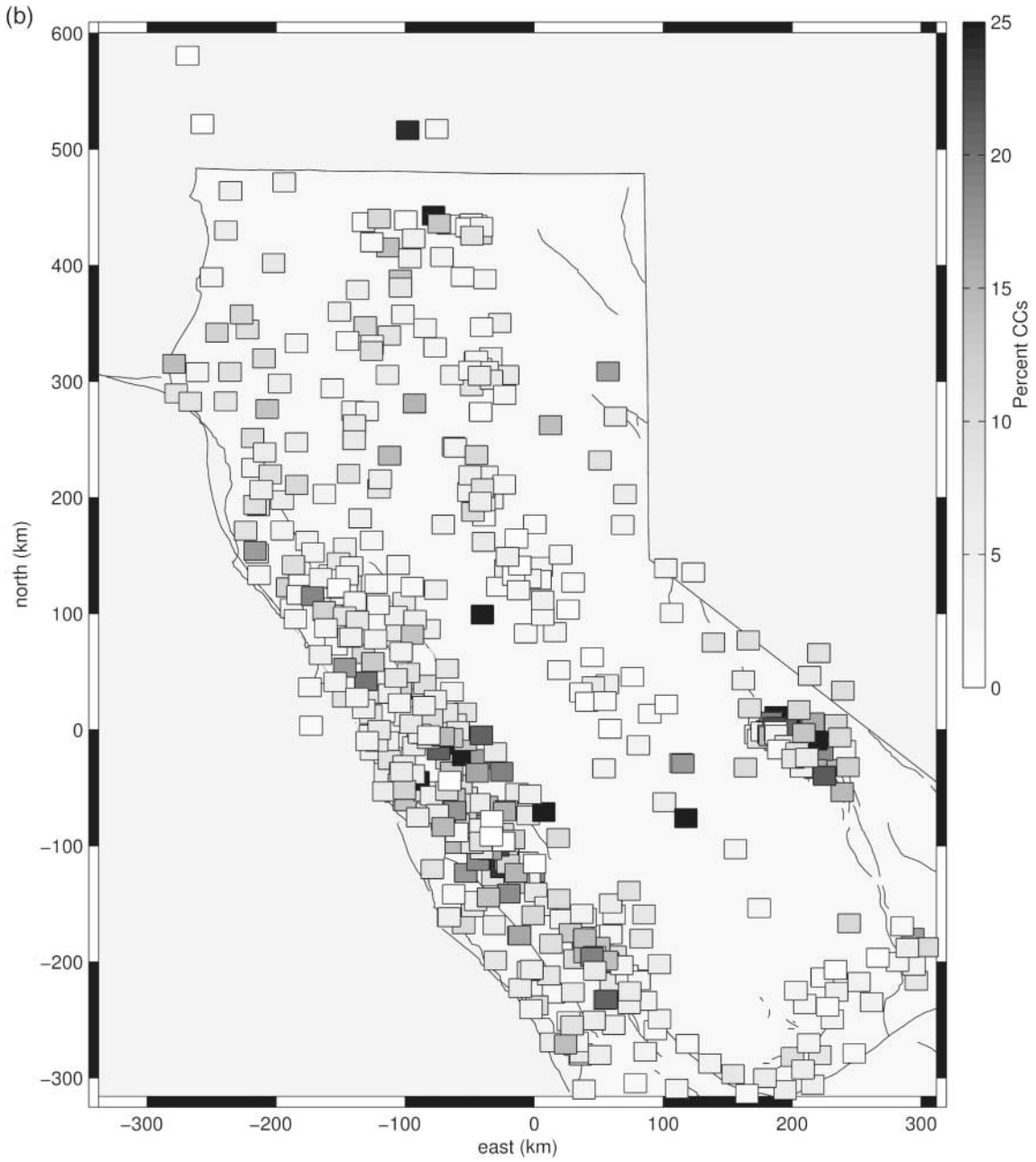

Figure 6. Continued.

1D layered velocity model, we are also able to increase the number of observations by about 30% compared to if we only used event pairs that had *P* picks for both events listed in the NCSN bulletin (see Table 2). Since there are virtually no *S*-wave picks in the phase data for the 225,000 events at the NCEDC, we use theoretical initial window alignments based on 1.732 times the *P*-wave travel time to perform cross correlations on windows containing *S*-wave energy and are able to obtain nearly the same number of *S*-wave observations as for *P* waves (Fig. 7b, Tables 1 and 2). (Note: *S*-wave correlations are measured on vertical components as well, since that is predominantly what is available.) Filtering from 1.5 to 15 Hz also increases the number of useful measurements by reducing long-period instrument noise and less similar high frequencies (Fig. 7b, Tables 1 and 2). Based on these results, we choose to filter all the seismograms from 1.5 to 15 Hz for the correlation database.

Discussion

The following figures illustrate the characteristics of the cross-correlation data and were used to set appropriate parameters for the processing. Figure 8a shows the contours of the distribution of CC versus interevent separation distance

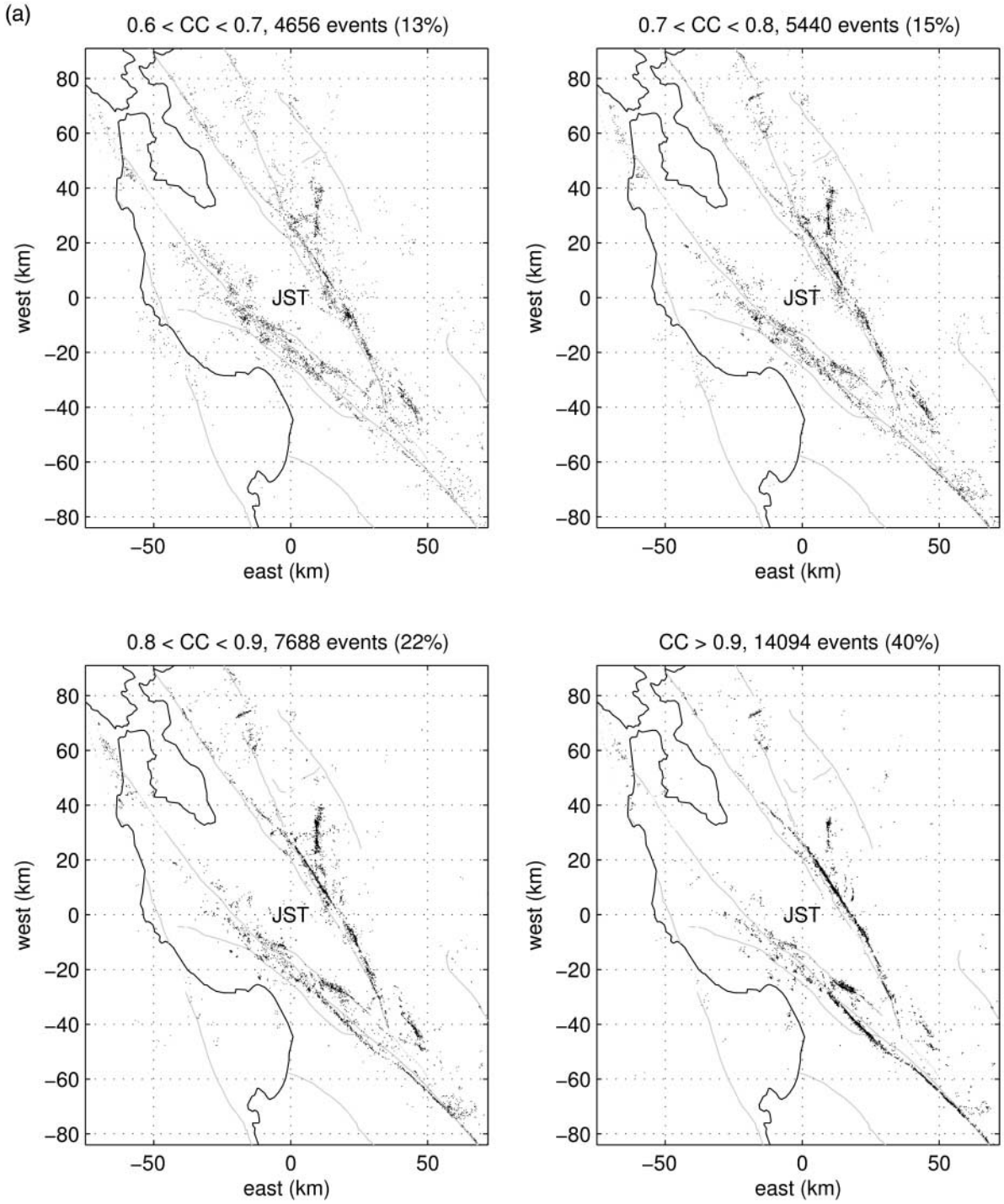

Figure 7. Cross-correlation measurements at station JST. (a) Double-difference locations computed using phase picks only shown for different CC intervals. (b) Histogram of events that correlate with at least one other events at the threshold indicated.

for station JST, at different confidence levels (a familiar plot) (e.g., Aster and Scott, 1993). For example, at 3-km interevent distance, 95% of the event pairs have CC $<$ 0.6 for this distance bin. CC values decrease as expected because of the breakdown in waveform similarity with increasing separation. We see that a maximum event separation of 5 km captures most of the useful cross-correlation measurements. By useful we mean that the events are close enough to produce adequately similar waveforms such that the differential travel times provide an improvement over phase picks. From this plot (and also from Fig. 7a for lower CC) we see event pairs with longer separation distances can be used, rather

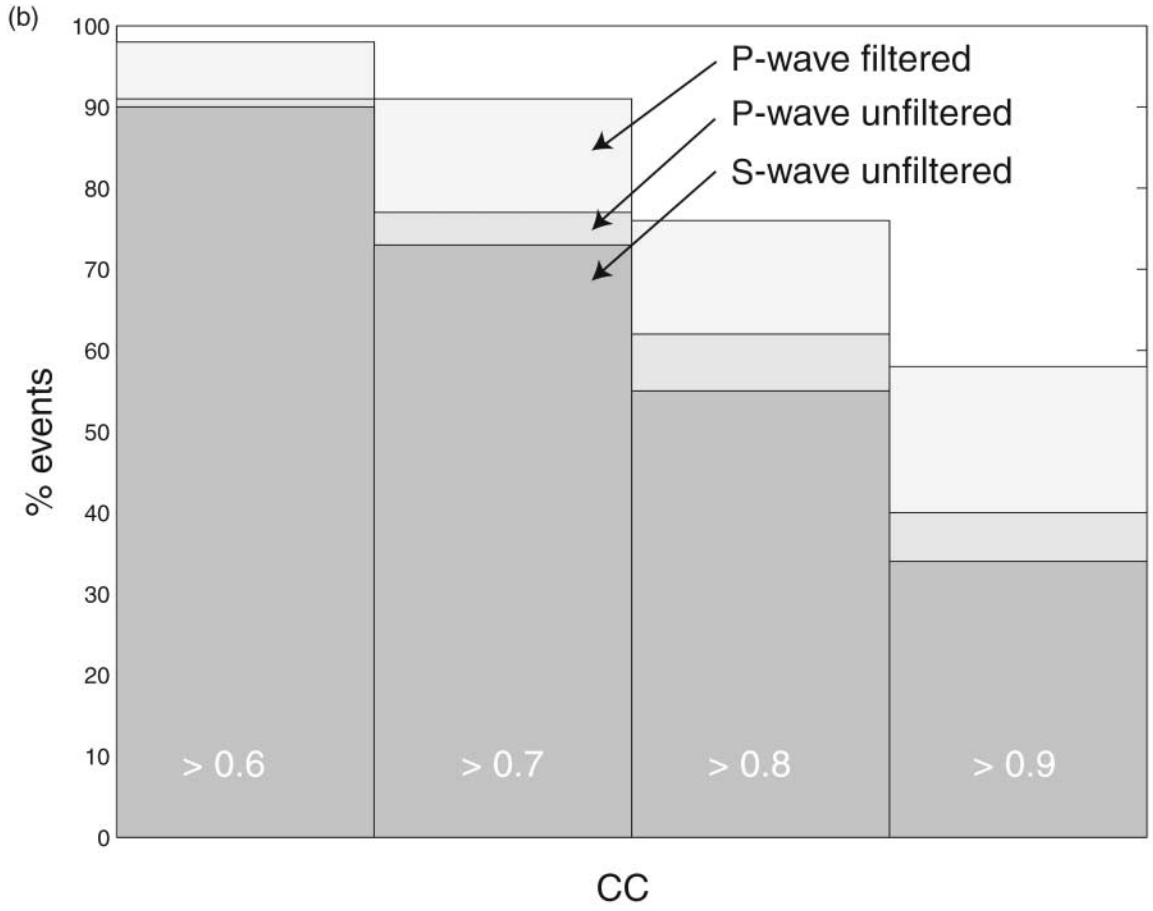

Figure 7. Continued.

than just repeating or colocated events. Note that 5 km is quite a bit longer than a quarter wavelength rule for the dominant frequencies in a 1.5- to 15-Hz band (Geller and Mueller, 1980), which for a P-wave velocity of 6 km/sec would be from 0.1 to 1 km. It was shown for a case example on the Calaveras fault that often longer interevent distances (up to 2 km) can be used for cross correlation (Schaff *et al.*, 2004) if the mechanisms are similar and path effects are consistent for a common station. If the crustal structure is relatively homogeneous, the waveforms are expected to be similar for greater interevent separation distances. Comparing plots for other stations showed similar behavior. Based on these examples and allowing for average uncertainties in the initial event locations, we chose an interevent separation distance threshold of 5 km for our cross-correlation analysis.

We explored the dependence of cross correlation on distance to the station. In theory, if the Earth were to act like a low-pass filter, CC would increase with increasing station distance, because high-frequency energy is more subject to scattering on small-scale heterogeneities. In a similar way we might expect that the ratio of interevent distance to station distance is important. For example, if two events are separated by 1 km and are recorded at two stations, one 10 km away and the other 100 km away, we would expect the correlation coefficient to be higher at the more distant station. An example of this behavior of increasing CC with station distance is observed at the Long Valley caldera station, MDR (Fig. 8b). It suggests that the crust in this region is perhaps uniformly fractured or has some other properties (e.g., attenuation) that cause it to act like a low-pass filter for the propagating seismic waves. This phenomenon, however, was not a widely observed at all the other stations. The breakdown between 20- to 30-km station distance may be due to the crossover distance from direct waves to refracted arrivals. After 30-km station distance, the trend is again seen

Table 1
Number of Correlated Events*

	CC ≥ Threshold			
Station (Phase)	0.6	0.7	0.8	0.9
JST (*P* wave)	32 K (91%)	27 K (77%)	22 K (62%)	14 K (40%)
MDR (*P* wave)	58 K (81%)	483 K (67%)	31 K (43%)	13 K (18%)
KBB (*P* wave)	14 K (78%)	10 K (57%)	6 K (36%)	3 K (19%)
JST (*S* wave)	31 K (90%)	25 K (73%)	19 K (55%)	12 K (34%)
JST (*P*-wave filtered)	34 K (98%)	32 K (91%)	27 K (76%)	20 K (58%)

*Numbers in parentheses are the percentages out of the total number of events.

to increase, although with more scatter, perhaps due to the larger area sampled and fewer events at these distances.

Another interesting aspect we explored with the new data is the degree to which crustal heterogeneity between source region and recording stations controls waveform similarity. We use a subset of about 1500 precisely located events along the Big Streak on the Parkfield section of the SAF (Waldhauser *et al.*, 2004). For each event pair/station configuration for which a cross-correlation coefficient of 0.7 or larger is obtained, we determine precise interevent distance and the azimuth between the direction of the event pair and the station that recorded both events. Figure 9 shows the variation of these CCs as a function of the event pair/station azimuth and different intervals of event separation. As ex-

Table 2
Number of Correlation Measurements*

	CC ≥ Threshold			
Station (Phase)	0.6	0.7	0.8	0.9
JST (*P* wave)	1.3 M (7%)	495 K (3%)	165 K (0.9%)	43 K (0.2%)
MDR (*P* wave)	5.1 M (5%)	1.5 M (1%)	355 K (0.3%)	29 K (0.03%)
KBB (*P* wave)	293 K (21%)	114 K (8%)	38 K (3%)	9 K (0.7%)
JST (*S* wave)	1.7 M (9%)	656 K (3%)	215 K (1%)	54 K (0.3%)
JST (theor. *P* wave)†	308 K (30%)	105 K (28%)	36 K (27%)	10 K (31%)
JST (*P*-wave filtered)	4.1 M (21%)	1.7 M (9%)	578 K (3%)	136 K (0.7%)

*Numbers in parentheses are the percentages out of the total number of correlations computed.

†Values in parentheses are the percent increases from the first JST row obtained by including initial theoretical alignments.

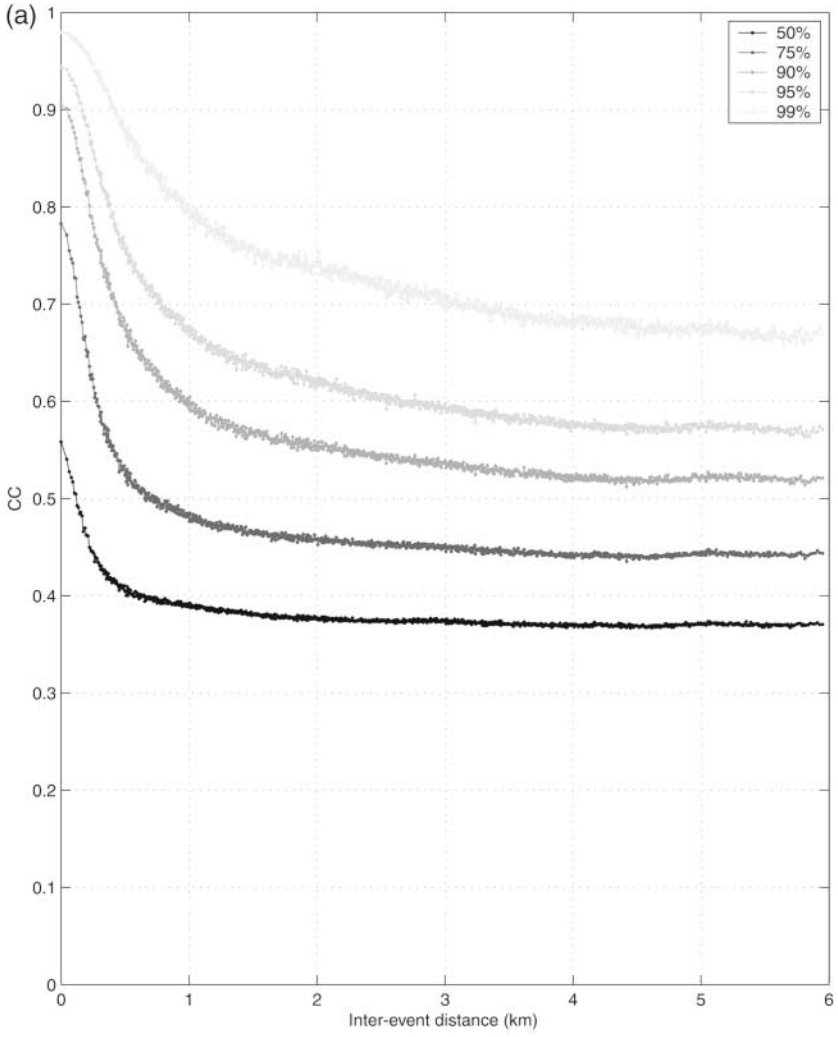

Figure 8. (a) Statistics of correlation coefficients for station JST showing the breakdown of waveform similarity with event separation distance. Contours show confidence levels defined in the legend. They are computed by dividing the *x*-axis into 1000 bins of equal number represented by each point (e.g., JST has 1900 observations per bin). (b) Distribution of CC versus station distance for Long Valley caldera station, MDR. (*continued*)

pected, for events that are colocated, the CCs are insensitive to variation in recording azimuth. With increasing recording azimuth, and within intervals of interevent distances, we observe a trend of CC decrease. This is because at zero azimuth the rays for two events travel a similar path outside the source region since the station is in direction of the event pair. At an azimuth of 90° ray paths are perpendicular to the relative position vector of the event pair, and thus travel through increasingly different media, compared to the case where the event separation and slowness vectors are parallel (0°).

The CC is not a fixed quantity for two seismograms. It varies with certain parameters such as filtering and window length. It is important to understand how CC depends on these factors because it is most often used to set thresholds and weights for data quality. For example, shorter window lengths will have higher CCs than longer window lengths for the same records. Even for the case of purely random noise waveforms (where CC approaches zero as window length goes towards infinity), as the window length approaches one sample, CC approaches unity. A simple synthetic test proves this result. The reason is that shorter windows appear less random because of the fewer data points and the issue of statistics on small sample sizes. Figure 10a shows the distribution of contours of CC computed for window lengths of 1 and 2 sec at station JST. Since most of the contours are above the $y = x$ line it verifies that CCs are higher for shorter window lengths. The 50% median contour line can be thought of as a useful way to map CC for one window length to another. For example, statistically speaking, a CC threshold of 0.7 for a window length of 1 sec is equivalent to a CC threshold of 0.6 for a window length of 2 sec. In this way a quantitative understanding of CC thresholds for different window lengths can be determined.

Using correlation-coefficient thresholds is currently the primary means for deciding what data to include for future location studies. We sought additional independent means to judge measurement quality and remove outliers. Computing correlations at two different window lengths provides two independent relative arrival time measurements that should agree for the same phase at the same station. Figure 10b shows the distribution of the absolute values of the difference in delay times for the two window lengths, abs(dt2 − dt1). For station JST, which has many similar

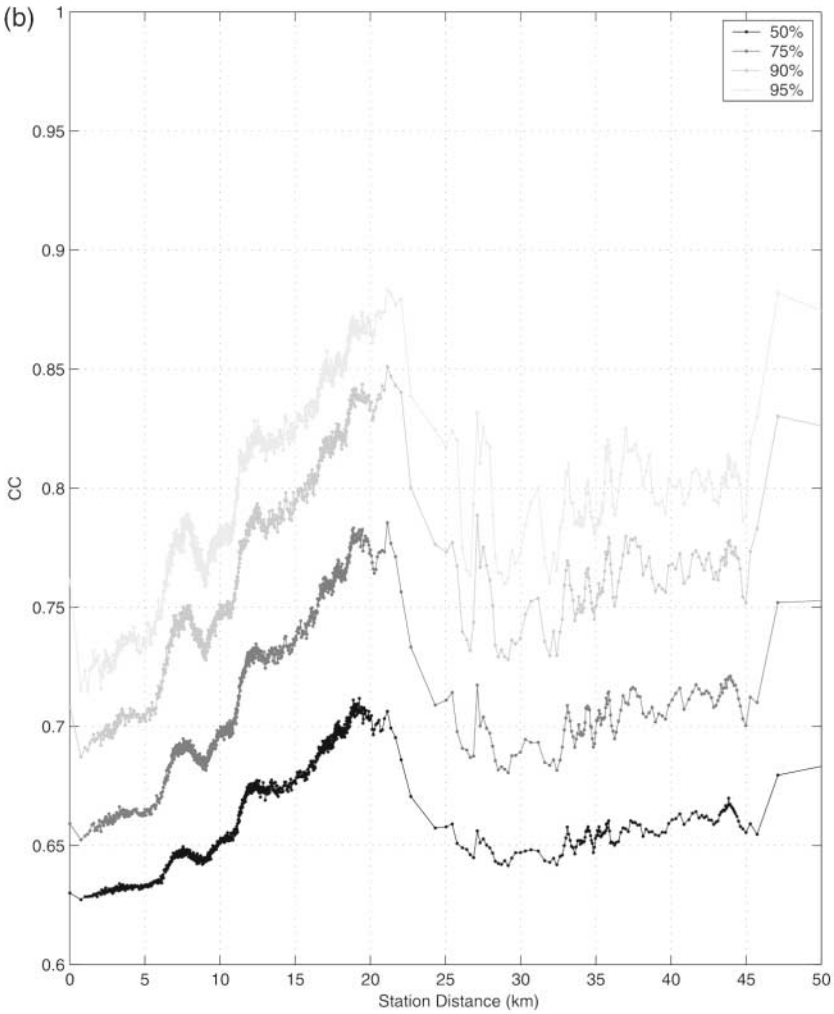

Figure 8. Continued.

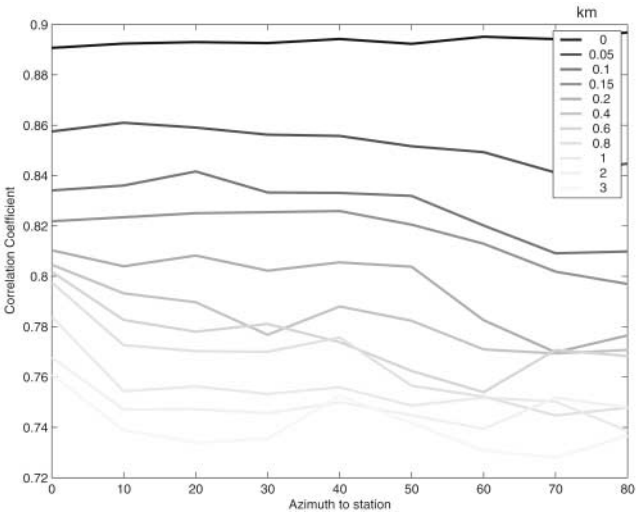

Figure 9. CCs as a function of recording azimuth (station relative to direction of event pair), for different intervals of interevent distances (km) shown by the gray scale in the legend. Data are from 1500 events along a shallow streak on the Parkfield section of the SAF.

events, the values agree to two samples (0.02 sec) or better for the 95% confidence interval up to CC = 0.6. Combined with CC thresholds this can be an additional way to remove measurement outliers. From such a procedure we were also able to determine that filtering can remove some large outliers associated with long-period instrument noise even though the CCs were high and therefore not excluded on that basis. Other ways of removing outliers and judging data quality are to use error estimates for the delay measurement itself (Aster and Rowe, 2000; Rowe *et al.*, 2002) and bispectrum verified delays (Du *et al.*, 2004). These procedures are, however, not feasible in the scope of this project owing to their low computational efficiency. Also, the bispectrum verified method, which cancels common noise (mostly of environmental nature) works predominately for events close in time (e.g., aftershocks), recorded during the same environmental conditions. Such noise we are able to filter out as they contain mostly high frequencies.

Conclusions

This study resulted in a wealth of cross-correlation and differential time information, consistently measured for all events digitally recorded by the NCSN between 1984 and 2003. Preliminary inspection of these data indicates its potential usefulness in a wide range of future research, including but not limited to regional-scale earthquake relocation studies and tomographic investigations, as well as characterization of crustal heterogeneity across northern California. Double-difference relocations of the NCSN catalog with only the phase data showed a substantially increased level of detail across most of the northern California region (see, e.g., Fig. 4, Fig. 7a), which can be significantly enhanced by incorporating the cross-correlation differential times presented here. To solve for the four unknowns of an earthquake hypocenter location, at least four stations are needed. This project had the surprising result that 95% of the seismic events with waveforms available in northern California had CC ≥ 0.7 at four or more stations with at least one other event. Therefore it is expected that the majority of the earthquake locations may be improved by the correlation measurements.

Implementation of a correlation detector as opposed to a correlation function may be partially responsible for the high percentage of similar event pairs discovered since it can recover delays of arbitrarily long offset without degradation in the measured correlation coefficient value. Another reason is that we chose generous interevent distance thresholds of 5 km, so as not to miss any potential similar event pairs due to initial location errors. Dependencies of correlation coefficient on interevent distance, station distance, and recording azimuth were observed as expected and varied quantitatively by station and region. In order to judge data quality and remove outliers, it was demonstrated that the correlation coefficient of one window length could be mapped to that of another window length to set appropriate thresholds and understand their significance. An alternate way of removing outliers was presented by measuring the differential travel times for the same phase using two different window lengths and assessing their agreement.

Some of the case studies in northern California that showed high percentages of similar events were on creeping sections of the San Andreas south of Loma Prieta (Rubin *et al.*, 1999) and at Parkfield (Waldhauser *et al.*, 2004). Another example is on the partially creeping Calaveras fault where 92% of the earthquakes had sufficient similarity for the locations to improve by one to two orders of magnitude (Schaff *et al.*, 2002). Our processing was uniformly computed across northern California allowing direct comparison for different tectonic areas of various complexity such as the SAF system, Long Valley caldera, Geysers geothermal field, and Mendocino triple junction. Especially at Long Valley and Geysers, which produced about half of the 225,000 events, it is surprising the high degree of correlation despite their complicated 3D structures (Fig. 5). The velocity structure and fault orientations vary strongly in both of these regions, suggesting that waveform similarity would not be as predominant as in creeping zones. The most likely explanation is the intense earthquake density for both areas. There is a high probability of at least one other earthquake occurring nearby over the 19-year period with a similar source mechanism. Similar efforts in southern California, also predominantly complex and 3D in nature, indicate that the locations of 65% of the events can be improved with waveform-based delay time measurements (Hauksson and Shearer, 2005).

The results presented here point to the value of maintaining a permanent network of stations over long periods of time. With larger archives there is a greater chance that

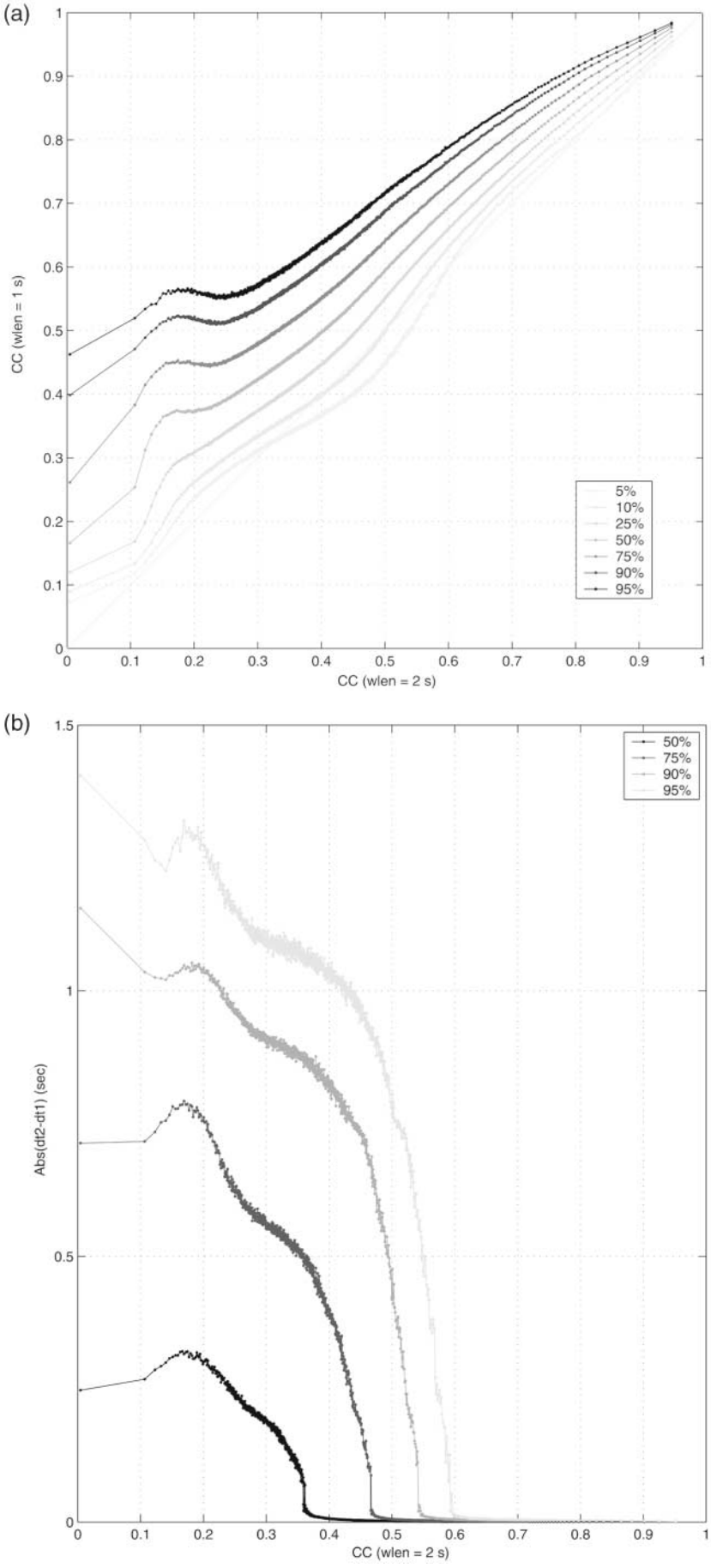

Figure 10. (a) Distribution of CC for 1-sec window lengths versus CC for 2-sec window lengths. Contours show confidence levels defined in the legend (similar description as for Figure 8). (b) Agreement of delay measurements (dt) for two different window lengths (1 and 2 sec). Distribution of absolute difference in the delay times versus CC (2-sec window lengths).

new events will match other events already recorded. We see that station density is also important since at least four stations are needed to locate the event and differences in the radiation pattern or weak signal-to-noise ratios may make some correlations unusable at certain stations (Fig. 6). The situation is expected to only improve over time as both earthquake density and station density increase offering greater possibilities for waveform cross-correlation differential travel times to supplement and improve one of the most fundamental datasets in seismology—arrival times.

Acknowledgments

We thank the network operators from the U.S. Geological Survey (USGS) and Berkeley Seismological Laboratory (BSL) for maintaining a high quality, long-running permanent short period network (NCSN). Special thanks goes to Doug Neuhauser, manager of the NCEDC, for extracting a copy of the database for us. We are grateful to Lamont's computer support, Doug Shearer and Bob Bookbinder, for additional help in this project. Support from Gus Correa and Marc Spiegelman was invaluable for assistance with the Linux cluster. Won-Young Kim aided in data handling operations. We thank Doug Dodge for help with the correlation algorithm. Comments from Associate Editor Charlotte Rowe, Peter Shearer, and an anonymous reviewer are appreciated. This work was supported by the USGS/NEHRP Grant 03HQGR0004. This is Contribution Number 6705 of the Lamont-Doherty Earth Observatory, Columbia University.

References

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.

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

Deichmann, N., and M. Garcia-Fernandez (1992). Rupture geometry from high-precision relative hypocentre locations of microearthquake ruptures, *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.

Du, W.-X., C. H. Thurber, and D. Eberhart-Phillips (2004). Earthquake relocation using cross-correlation time delay estimates verified with the bispectrum method, *Bull. Seism. Soc. Am.* **94,** 856–866.

Fréchet, J. (1985). Sismogenè et doublets sismiques, *Thèse d'Etat*, Université Scientifique et Medicale de Grenoble, 206 pp.

Frémont, M.-J., and S. D. Malone (1987). High precision relative locations of earthquakes at Mount St. Helens, Washington, *J. Geophys. Res.* **92,** 10,233–10,236.

Geller, R. J., and C. S. Mueller (1980). Four similar earthquakes in central California, *Geophys. Res. Lett.* **7,** 821–824.

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.

Hauksson, E., and P. Shearer (2005). Southern California hypocenter relocation with waveform cross-correlation, part 1: Results using the double-difference method, *Bull. Seism. Soc. Am.* **95,** 896–903.

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

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

Moriya, H., H. Niitsuma, and R. Baria (2003). Multiplet-clustering analysis reveals structural details within the seismic cloud at the Soultz Geothermal Field, France, *Bull. Seism. Soc. Am.* **93,** 1606–1620.

Nadeau, R. M., W. Foxall, and T. V. McEvilly (1995). Clustering and periodic recurrence of microseismicities on the San Andreas fault at Parkfield, California, *Science* **267,** 503–507.

Neuhauser, D. S., B. Bogaert, and B. Romanowitz (1994). Data access of Northern California seismic data from the Northern California Earthquake Data Center (abstract), *EOS* **75,** 429.

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

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

Rowe, C. A., R. C. Aster, B. Borchers, and C. J. Young (2002). An automatic, adaptive algorithm for refining phase picks in large seismic data sets, *Bull. Seism. Soc. Am.* **92,** 1660–1674.

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

Schaff, D. P., and G. C. Beroza (2004). Coseismic and postseismic velocity changes measured by repeating earthquakes, *J. Geophys. Res.* **109,** B10302, doi 10.1029/2004JB003011.

Schaff, D. P., G. H. R. Bokelmann, G. C. Beroza, F. Waldhauser, and W. L. Ellsworth (2002). High resolution image of Calaveras Fault seismicity, *J. Geophys. Res.* **107,** 2186, doi 10.1029/2001JB000633.

Schaff, D. P., G. H. R. Bokelmann, W. L. Ellsworth, E. Zanzerkia, F. Waldhauser, and G. C. Beroza (2004). Optimizing correlation techniques for improved earthquake location, *Bull. Seism. Soc. Am.* **94,** 705–721.

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., E. Hauksson, and G. Lin (2005). Southern California hypocenter relocation with waveform cross-correlation, part 2: Results using source-specific station terms and cluster analysis, *Bull. Seism. Soc. Am.* **95,** 904–915.

Waldhauser, F. (2001). HypoDD: a computer program to compute double-difference hypocenter locations, *U.S. Geol. Surv. Open-File Rept. 01-113*, 25 pp.

Waldhauser, F., and W. L. Ellsworth (2000). A double-difference earthquake location algorithm: method and application to the northern Hayward Fault, California, *Bull. Seism. Soc. Am.* **90,** 1353–1368.

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

Waldhauser, F., W. L. Ellsworth, D. P. Schaff, and A. Cole (2004). Streaks, multiplets, and holes: high-resolution spatio-temporal behavior of Parkfield seismicity, *Geophys. Res. Lett.* **31,** L18608, doi 10.1029/2004GL020649.

Lamont-Doherty Earth Observatory
Columbia University
61 Route 9W
Palisades, New York 10964

Manuscript received 23 November 2004.