

Improved Differential Travel Time Measurements and a Search for Repeating Events at the Northern California Seismic Network

Grant 03HQGR0004

Felix Waldhauser

Lamont-Doherty Earth Observatory of Columbia University, Palisades, NY 10964

Tel: (845) 365 8538; Fax: (845) 365 8150; felixw@ldeo.columbia.edu

URL: <http://www.ldeo.columbia.edu/~felixw>

NEHRP Element I, Products for Earthquake Loss Reduction

Keywords: Seismology, Source characteristics, Database

Annual Project Summary for FY 2003

This report covers the activities between December 1, 2002 (start date of the project) until September 30, 2003 of this one-year project. The work described in this report is being undertaken by the principle investigator Felix Waldhauser and, for the larger part, by David Schaff (post doctoral research scientist). The research includes the collection and reformatting of all digital waveform data recorded by the Northern California Seismic Network (NCSN), the development of computational tools to measure differential travel times by waveform cross correlation on a massive scale, the application of these tools to the NCSN waveforms, and a search for the existence of repeating events.

1. Investigations undertaken

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 travel time curves. But there is an error associated with each measurement. Average pick errors for Northern California Seismic Network (NCSN) phase data are on the order of 0.1 sec. These errors map into significant scatter in the earthquake locations and reduce the resolution of tomographic inversions.

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. This directly translates into substantially more precise event locations. Under this grant we are developing a data base of cross correlation information and differential travel time measurements for all similar events across the Northern California Seismic Network (NCSN), using the digital waveforms stored at the Northern California Earthquake Data Center (NCEDC) for events between 1984 and 2001. We are currently in the process of performing 2.5 billion correlation measurements on 280,000 events in Northern California from 1984 to present. The waveforms comprise the complete digital archive at the NCEDC recorded by 900 short period vertical component stations totaling 225 GB of data. The analysis will be performed uniformly over the entire Northern California region. In this way, different areas may be compared objectively such as NEHRP priority faults: Hayward, Calaveras, and the San Andreas, as well as other hazard areas like Long Valley Caldera.

In FY 2003 we have performed the following steps necessary to complete the project:

- *Collection, local storage, and reformatting of the entire NCSN waveform data base.*

Doug Neuhauser at the NCEDC kindly extracted all of the 225 GB of waveform data to 10 DLT tapes. The data is stored in a compressed binary CUSP format which we convert to SAC file format for processing. The seismograms are also arranged by event in calendar time. The correlation processing reads all the events recorded at one station at a time, performs the correlation measurements for the desired pairs of events, and then proceeds to the next station. Therefore, the seismograms need to be reorganized from a calendar ordering to a station ordering. To accomplish this, it is most expedient to have enough disk space to accommodate all 225 GB at one time. For this we purchased two internal 120 GB hard drives from the funds of this project. The seismograms are uncompressed, preprocessed with travel time information for P- and S-waves being updated to the event headers, and then recompressed. These operations are performed for 15 million SAC files or seismograms. Data transfer rates from disk and across our network are on the order of 1 MB/sec. This amounts to about 3 days of computer time for each operation if uninterrupted — copying the data from the DLT drive, uncompressing, converting from CUSP to SAC, recompressing, and reorganizing into station subdirectories. The DLT drive, however, can only be manually operated and the tapes must be changed after each extraction is complete, increasing the number of days required. The total amount of time involved for the data manipulation and reformatting stage was about 2 months, accounting for the computer time and also the time needed to write the data handling routines and test the integrity of the transfer and conversion.

- *Development of computational tools to perform cross correlation measurements on a massive scale.*

Our earlier work used cross correlation routines that were designed to satisfy the memory and speed requirements for processing on the order of 10K events (Schaff et al., 2002; Schaff et al., 2003). Now with the task of processing over an order of magnitude more data, these routines needed to be modified to efficiently process larger numbers of events recorded by a single station. The initial correlation program used FORTRAN subroutines for the number crunching and MATLAB to facilitate the bookkeeping. To improve both the memory and speed, we have converted the whole program to FORTRAN and added some new features. Resampling and filtering can be performed on-the-fly internally within core memory. Also, a toggle feature for byte swapping has been included so that data can be analyzed on both Sun and Linux platforms. Depending on the window lengths and the lags searched over, we are able to perform about 10 million correlation measurements per hour.

- *Initial test runs and performance evaluation for 14 NCSN stations.*

Before embarking on massive processing of all the NCSN waveform data, we executed a battery of tests to evaluate the performance of the correlation method and judge which parameters give the best results uniformly across diverse tectonic settings. In order to debug the new code, we also compared it to the earlier version to make sure the results were reproducible. We explored the use of filtering and resampling to obtain more usable correlation measurements as well as the incorporation of theoretical P- and S-wave initial window alignments, if no phase picks were available. We experimented with different event separation distance cut-offs and using improved double-difference locations from phase data to determine the initial

pairs. We also examined the effects of different window lengths on the correlation coefficients (CC) and the robustness of the delay measurements.

Our first goal for this project is to produce the correlation database and then make it available to the greater seismological community because of the potential benefit improved differential travel times have for many diverse areas of research. The second goal is to identify and analyze repeating events throughout Northern California which may directly impact several of the NEHRP research priorities for the region and help to refine the urban hazard maps produced by the USGS.

2. Results

To improve the accuracy of inter-event distances from which we determine pairs of events for correlation measurements we have relocated about 225,000 events using the double-difference method together with about 5 million NCSN P-phase picks. The mean shift between routine NCSN locations and DD relocations is about 300 and 500 m in the horizontal and vertical direction, respectively. The relocated seismicity shows a substantial increased level of detail across most of the Northern California region, which can be significantly enhanced by incorporating the cross correlation differential time measurements, once they are available.

We have experimented with a correlation detector which is able to recover lags greater than half the window length. This is a new feature and different than the correlation function which was applied in our earlier work. Figure 1 shows examples of automatically determined P-wave arrival time adjustments of similar events observed at stations JST. These P-wave trains have $CC > 0.9$ and adjustments > 0.9 sec for window lengths of 1 sec. All of these event pairs had at least one theoretical initial window alignment, which is the reason for the large adjustments.

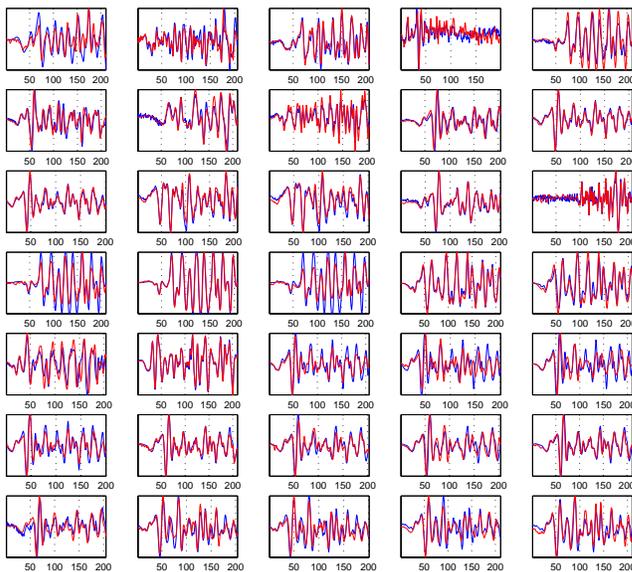


Figure 1 Aligned P-waves for several pairs of events (blue and red overlaying seismograms) obtained from a correlation detector. All adjustments are > 0.9 sec which is more than half the window length of 1 sec. The P-wave trains are very similar with $CC > 0.9$. X-axes are in samples (delta = 0.01 sec.).

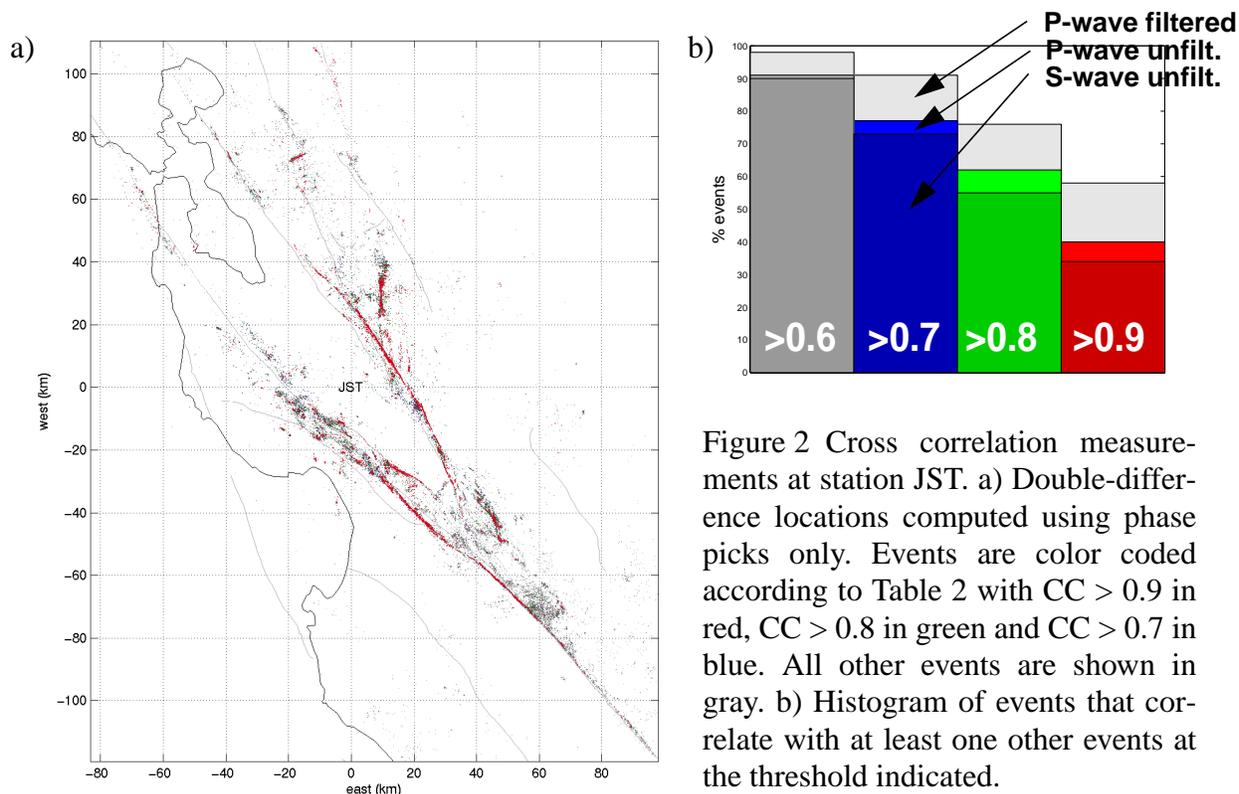


Figure 2 Cross correlation measurements at station JST. a) Double-difference locations computed using phase picks only. Events are color coded according to Table 2 with $CC > 0.9$ in red, $CC > 0.8$ in green and $CC > 0.7$ in blue. All other events are shown in gray. b) Histogram of events that correlate with at least one other events at the threshold indicated.

Event pair cross correlations are computed at a common station. At station JST, which includes 35,000 events from the San Andreas Fault system, 40% of the events have at least one other event with cross correlation coefficients (CC) greater than 0.9 (62% for $CC > 0.8$, 77% for $CC > 0.7$) (Figure 2). The percentages of similar events observed at station JST are surprisingly high, but they include known areas of repeating events on the Calaveras and San Andreas Faults. For a station in Long Valley Caldera (MDR) recording 72,000 events, the distribution is 18% for $CC > 0.9$, 43% for $CC > 0.8$, and 67% for $CC > 0.7$. A station including 20,000 events in the different tectonic settings of Mendocino Triple Junction and Geysers Geothermal Fields yields correlation measurements where 16% of the events have at least one $CC > 0.9$, 32% have $CC > 0.8$, and 49% have $CC > 0.7$. The lower numbers of correlated events observed at the latter two stations most likely reflect the different 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 2 that a large percentage of the events correlate well across a variety of tectonic regions. The correlations are expected to provide improved relative arrival times for performing double-difference locations for much of the seismicity.

There are no S-wave picks in the phase data for the 280,000 events at the NCEDC. By using theoretical initial window alignments based on 1.732 times the P-wave travel time and performing cross correlations on windows containing S-wave energy, we are able to obtain nearly the same number of S-wave observations as for P-waves (Figure 2b, Tables 1 & 2). We are therefore able to nearly double the number of observations that can be used for the location by including S-waves

that haven't been picked and also benefit from the extra constraint they provide on depth. Filtering is also seen to increase the number of observations that can be used for location (Figure 2b, Tables 1 & 2). 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 based on raytracing through a 1D 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 (Table 1).

Table 1: Number of Correlation Measurements

| station (phase) | CC > thresh | | | |
|-----------------------|-------------|-------------|--------------|--------------|
| | 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%) |

Table 2: Number of Events

| station (phase) | CC > thresh | | | |
|-----------------------|-------------|-------------|------------|------------|
| | 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%) |

We also examined the statistics of various parameters involved with the correlation to help gauge appropriate thresholds, values, and judge quality in order to remove outliers before the location. Since the number of correlation measurements goes like the number of events squared, it is impractical to do all possible correlations. In addition, from a quarter wavelength rule we don't expect events separated by great distances to correlate well (Geller and Mueller, 1980). Figure 3a shows the contours of the distribution of CC vs. inter-event separation distance for station JST. It decreases as expected because of the breakdown in waveform similarity with increasing separation. The different confidence levels are shown 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 obs per bin). From figures like this we are able to determine that event separations of 5 km and less should probably capture most of the useful cross correlation measurements.

Using correlation coefficient thresholds is currently the primary means for deciding what data to include for the location. 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 3b shows the distribution of the difference in absolute adjustments for two window lengths, $\text{abs}(\text{dt2}-\text{dt1})$. For station JST, which has lots of similar events, the values agree to two samples (0.02 sec) or better all the way out to $\text{CC} = 0.6$. Combined with CC thresholds this can be an additional way to remove outlier measurements. 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 correlation coefficients were high and therefore not excluded on that basis.

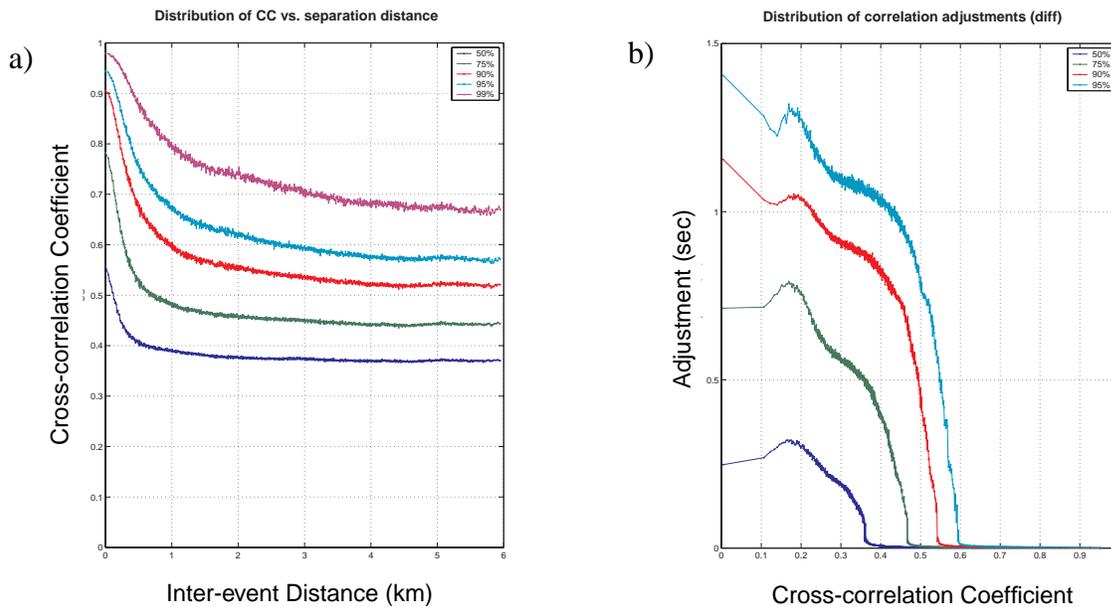


Figure 3 Statistics of correlation measurements for station JST. (a) Breakdown of similarity with event separation distance. (b) Agreement of delay measurements for two different window lengths.

We extended our tests for an additional 11 stations to evaluate performance across all of Northern California and further aid parameter selection (Figure 4 and Table 3). The results are quite encouraging. With only 11 stations there are 125,675 events that correlate with at least one other event at the 0.7 level. This suggests that nearly half of the events in the catalog (45% of 280,000) show potential for having enough similarity to provide suitable cross correlation measurements for improving the double-difference locations obtained with just phase data alone. At the 0.9 level, 61,435 events or $\sim 25\%$ of the catalog show a high degree of similarity with at least one other event. When the correlations are computed for all the stations, we expect these percentages to rise and subsequently to be able to obtain high resolution location estimates for much of Northern California seismicity. We are in the last stages of porting the programs and waveform data over to a 32 node Linux cluster to complete the measurements for the remaining stations. It is estimated, with the dedicated resources of this cluster, that a single run over the whole archive will take approximately 5 days. Following that we will begin to mine the results for repeating earthquakes.

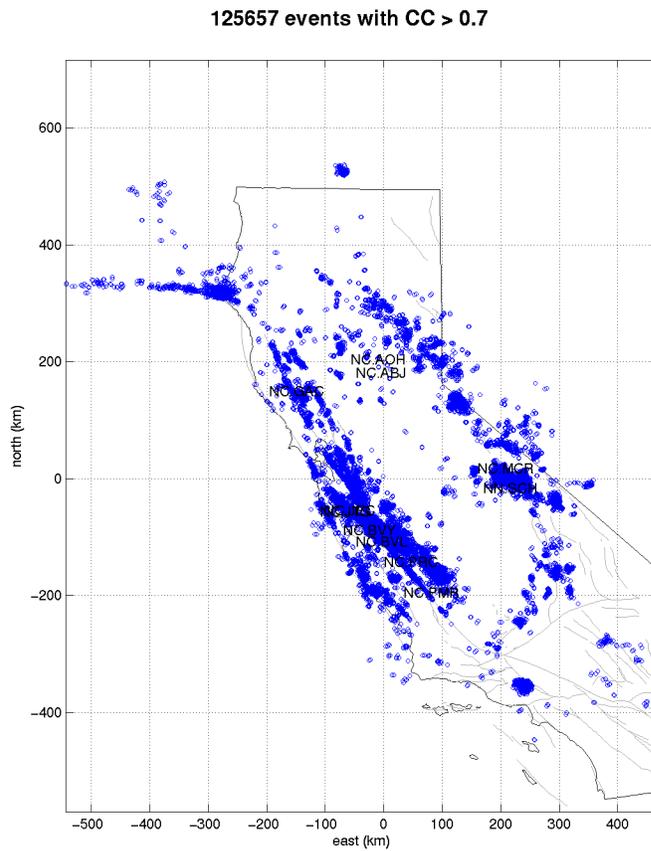


Figure 4 Results for 11 additional test stations. Events plotted (blue circles) have a correlation coefficient of 0.7 or greater with at least one other event at any of the 11 stations. There is potential that over half of the seismicity in Northern California may have useful correlation differential travel time data that can improve the event locations.

Table 3: Number of events for 11 additional test stations

| station | CC > thresh | | | |
|---------|-------------|------------|-------------|------------|
| | 0.6 | 0.7 | 0.8 | 0.9 |
| NC.ABJ | 13 K (66%) | 6 K (31%) | 2 K (10%) | 507 (3%) |
| NC.GAC | 16 K (86%) | 14 K (72%) | 8 K (43%) | 3 K (15%) |
| NC.PRC | 25 K (88%) | 19 K (66%) | 10 K (36%) | 5 K (19%) |
| NC.PMR | 19 K (79%) | 12 K (53%) | 6 K (27%) | 3 K (13%) |
| NC.MCR | 22 K (98%) | 21 K (93%) | 18 K (81%) | 13 K (57%) |
| NN.SCH | 22 K (94%) | 20 K (84%) | 12 K (53%) | 6 K (23%) |
| NC.AOH | 26 K (81%) | 17 K (53%) | 5 K (16%) | 731 (2%) |
| NC.BVL | 46 K (96%) | 42 K (88%) | 35 K (72.%) | 23 K (48%) |
| NC.BVY | 44 K (93%) | 39 K (83%) | 32 K (67%) | 22 K (47%) |
| NC.JEC | 43 K (91%) | 35 K (74%) | 21 K (44%) | 12 K (25%) |
| NC.JTG | 41 K (89%) | 32 K (69%) | 18 K (39%) | 10 K (21%) |

Improved Differential Travel Time Measurements and a Search for Repeating Events at the Northern California Seismic Network

Grant 03HQGR0004

Felix Waldhauser

Lamont-Doherty Earth Observatory of Columbia University, Palisades, NY 10964

Tel: (845) 365 8538; Fax: (845) 365 8150; felixw@ldeo.columbia.edu

URL: <http://www.ldeo.columbia.edu/~felixw>

NEHRP Topical Area 2. Northern California (NC).

Keywords: Seismology, Source characteristics, Database

3. Non-technical Summary

We compute differential travel times for similar earthquakes observed at common stations across Northern California. This database has the potential to improve by a factor of ten the precision of a fundamental measurement in seismology, which then can be used to better locate earthquakes and solve for the velocity structure in the crust, for example. Our initial analysis will mine this database for repeating events — earthquakes that occur close to one another and which have similar focal mechanisms resulting in nearly identical waveforms. Repeating events are finding several applications that may eventually help to reduce earthquake hazard in Northern California.

4. Reports published (related to this project)

Schaff, D. and F. Waldhauser, Progress in massive waveform cross correlation and wide area event relocation in Northern California, Proceedings and Abstracts, Volume XIII, Southern California Earthquake Center Annual Meeting, Oxnard, CA, Sept. 7-11, 2003.

Waldhauser, F. and D. Schaff, Cross-correlation and double-difference relocation in Northern California, presented at the SCEC Workshop on Converting Advances in Seismology into Earthquake Science, Caltech, Pasadena, September 22-23, 2003.

Schaff, D., F. Waldhauser, and P.G. Richards, Applying Massive Waveform Cross Correlation and Double-Difference Location to Northern California and China, abstract submitted to the 2003 Fall AGU meeting in San Francisco.

5. References

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

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. Zankerka, F. Waldhauser, and G. C. Beroza. Optimizing correlation techniques for improved earthquake location, *Bull. Seism. Soc. Am.*, in press.

6. Available Data

The NCSN waveform data used in this project is freely available at the NCEDC at UC Berkeley. Since a fair amount of effort is required to change the data from an event (calendar time) ordering scheme to a station based ordering, it is possible that we could make the reorganized 225 GB dataset available to interested researchers or even to the data center itself. Upon completion of this project it is expected that a waveform cross correlation database will be available for all the digital NCSN stations from 1984 to present. Due to funding level constraints, it was agreed in the revised budget for this project that the database would be released to other researchers interested in collaboration and analyzing the data for joint scientific inquiry. It is our goal to then soon afterward, make the database openly and publicly available because of the large potential benefit to the greater geophysical community.

For more information on data availability, contact:

Dr. David Schaff
Lamont-Doherty Earth Observatory
61 Route 9W
Palisades, NY 10964
dschaff@ldeo.columbia.edu