

# EVALUATION OF CROSS-CORRELATION METHODS ON A MASSIVE SCALE FOR ACCURATE RELOCATION OF SEISMIC EVENTS

Won-Young Kim, David P. Schaff, Jian Zhang, Felix Waldhauser, and Paul G. Richards

Columbia University

Sponsored by Air Force Research Laboratory

Contract No. F19628-03-C-0129

## **ABSTRACT**

Lamont-Doherty Earth Observatory (LDEO) at Columbia University is evaluating a method of locating seismic sources (earthquakes, explosions) based on the use of waveform cross-correlation (WCC) measurements instead of using the conventional measurements of seismic wave arrival time (phase picks). WCC measurements have been demonstrated to be 10 to 100 times more accurate, where they can be used. The principal issue we are exploring is the extent to which a significant fraction of seismicity can be located using WCC measurements. We have organized the work into two projects: (1) the application of WCC methods to relocate earthquakes and explosions in China and neighboring regions, and (2) the application of WCC methods to relocate earthquakes in parts of North America.

Using a very sparse network it was discovered that about 10% of earthquakes in and near China between 1985 and 2000 with  $M \geq 3$  were repeating events that generated essentially the same signals from sources which could not be more than 1 km from each other. The estimated location precision is a few 100 m. A specialized case study for the 1999 sequence of events in Xiuyan, China, found that WCC on Lg-waves combined with the double-difference technique significantly improved the epicentral locations. Several reflectivity synthetic experiments have been conducted to understand how depth, distance to the station, and mechanism influence the similarity of the Lg waveforms. In the synthetic Lg waveforms there is greater sensitivity to depth, than to epicentral distance.

In North America, we are studying the Charlevoix region in eastern Canada and the New Madrid seismic zone in the central United States. Datasets have been assembled from scratch working in conjunction with regional network operators. For Charlevoix courtesy, of the Geological Survey of Canada, we now have 2,472 events with corresponding catalog, phase, and waveform data. For New Madrid, two datasets have been acquired. The first is from the PANDA deployment between 1989 and 1992, which consists of 884 events with bulletin and waveform information. The second is from the Center for Earthquake Research and Information (CERI) network, operated by Memphis State University, which we currently have waveform data from 1995-2003, and catalog and phase data from 2000-2003 for 680 events, with catalog and phase data from 1995-1999 expected shortly. Preliminary WCC results of the PANDA network indicate that 68% (597 out of 884 events) correlate with cross-correlation coefficients above 0.7 at four or more stations. Four stations are the minimum required to obtain a location estimate. Both P- and S-waves are correlated on all three components. The window lengths are 1 s and the lags searched over are also 1 s. It appears in a few examples that similar correlations are possible over 1 to 10 kilometer inter-event separation distances due to a site resonance from soft sediments underneath certain stations. Subsequent work will continue the correlation analysis for the other regions and then perform locations using the double-difference algorithm.

## **OBJECTIVES**

To evaluate a method of locating seismic sources (earthquakes, explosions) that is based on use of waveform cross-correlation (WCC) measurements instead of using the conventional measurements of seismic wave arrival time. WCC measurements are ten or a hundred times more accurate, where they can be used. The principal issue we shall explore is the extent to which a significant fraction of seismicity can be located using WCC measurements.

## **RESEARCH ACCOMPLISHED**

We have organized the work into two Projects. Project 1 concerns application of WCC methods to relocate earthquakes and explosions in China and neighboring regions. Project 2 concerns application of WCC methods to relocate earthquakes in parts of North America.

### **Project 1**

For Project 1, we studied 14,000 earthquakes in China for the years 1985 to 2000, and in particular 130,000 regional seismograms recorded at a sparse network of stations up to 20 degrees distance. We made the surprising discovery that in some cases the complex, highly scattered Lg-wave is remarkably similar for clusters of events. We analyzed in detail a subset of 28 events out of 90 from the 1999 Xiuyan sequence associated with a damaging earthquake that was successfully predicted in Liaoning Province, northeast China. Excellent relative locations could be obtained using only four or five stations 500 to 1000 km away. This approach was expanded to cover all of China and surrounding regions, and we found that about 10% of seismic events in and near China for the fifteen-year period were repeating events not more than about 1 km from each other. Specifically we found a set of 1301 seismic events with the property that any one of them wrote regional seismograms (in many cases lasting for hundreds of seconds, and using a band from 0.5 to 5 Hz) almost exactly like those from at least one other event. Among these events, 950 waveform doublets could be identified, which can be used to evaluate the precision of phase picks and standard catalogs. This work has been written up and accepted for publication in two papers (Schaff and Richards, 2004a,b).

We continued a detailed study of the Xiuyan sequence of earthquakes in Liaoning Province, China. Specifically, we used a reflectivity code to generate synthetic regional waves in the passband 0.5 to 5 Hz, at a distance of 750 km (corresponding to the distance at which we have data), and found that the synthetics are far more sensitive to changes in source depth than they are to changes in epicentral distance. This is of interest in the context of interpreting event clusters that have approximately the same epicenter centroid, yet have different waveforms. We speculate that in this case the waveforms are different and do not cross-correlate well between clusters, because the clusters are centered at different depths.

### **Project 2**

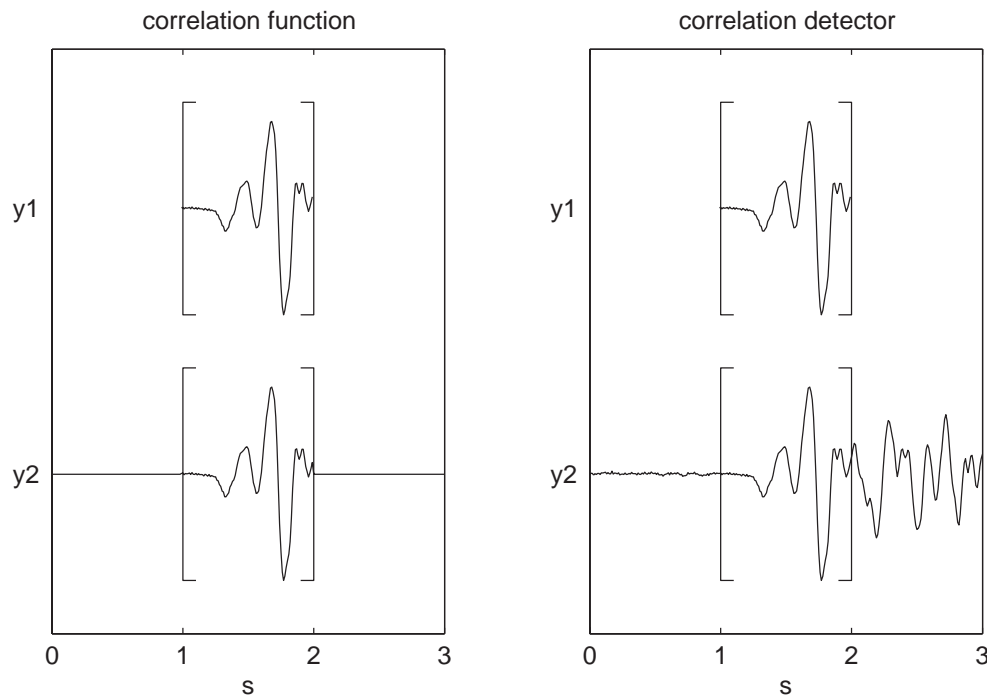
For Project 2, we are carrying out detailed studies of the seismicity of two regions in North America, namely New Madrid in the Central United States, and Charlevoix, East Canada.

#### *Data Compilation*

For New Madrid, three datasets have been acquired. The first is from the PANDA (Portable Array for Numerical Data Acquisition) network deployment from Oct. 1989 - Aug. 1991, which consists of 884 events with bulletin and waveform information. The second is from the network operated by CERl (Center for Earthquake Research and Information, University of Memphis) and we currently have waveform data from 1995 to 2003 and catalog and phase data from 2000 to 2003 for 680 events, with catalog and phase data from 1995 through 1999 expected shortly. The third data set is the Central Mississippi Valley seismic bulletin data from St. Louis University, which consists of over 60,000 phase picks from about 4,000 earthquakes that occurred during 1974-1998. For the Charlevoix seismic zone, courtesy of the Geological Survey of Canada, we now have 2,472 events with corresponding catalog, phase, and digital waveform data.

## Technique

We have experimented with a correlation detector that is able to recover lags greater than half the window length. This is a new feature and different from the correlation function which was applied in our earlier work. 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 (Figure 1). An equivalent result is obtained if the cross-correlation is computed in the frequency 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 that beyond this point the percentage of similar energy in the two windows is less than 50%. A related effect is that the cross correlation coefficient 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 (Figure 1). 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.



**Figure 1. Illustration of the difference between a correlation function and a correlation detector. Window is fixed for seismogram y1. A sliding window for seismogram y2 runs over zeros padding for the correlation function, and over data for the detector.**

Most implementations of cross correlation in the seismological literature we know of use a correlation function or the cross spectral technique, and are therefore limited by these fixed window length/finite duration records. 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 mispicks each by 0.5 s the total offset could range up to 1 s. A correlation function using 1 s window lengths would not be able to align these offsets.

Event pairs were selected for cross correlation computations based on distance thresholds since waveform similarity is known to decrease with increasing inter-event distance (e.g. Geller and Mueller, 1980). All event pairs with horizontal separations less than 5 km and vertical separations less than 10 km were processed. The input parameters used were window lengths of 1 s, lags searched over 1 s, and band-pass filtering from 1 to 15 Hz. All channels end components were processed where available. Measurements were made for both P- and S-waves, and phase picks listed in the bulletins were used to align windows initially. In cases where no phase picks were present, theoretical

travel times were computed using ray tracing through a 1D layered velocity model. If a P-wave pick existed but no S-wave pick existed, we used 1.732 times the P-wave travel time. Computations were performed on a Sun Blade 100 and processing time averaged on the order of a few hours.

*New Madrid Seismic Zone (PANDA deployment)*

Results for the 884 events recorded at the PANDA network (Figure 2) are summarized in Table 1 for correlation measurements of pairs of events that had cross-correlation coefficients  $CC > 0.7$  at four or more common stations. This table indicates that the location of 597 events or (68%) can be improved with the more accurate differential times from waveform cross correlation. If an event pair meets the criterion of  $CC > 0.7$  at four or more stations it is also an indication of the robustness of the measurements and quality of the data since there is less of a chance of outliers. Event pairs that have  $CC > 0.7$  at only one station may be more suspect as potential outliers. It may however be due to weak signal-to-noise ratios, stations coverage, or radiation patterns. Therefore in a joint inversion for all the events using the double-difference method (Waldhauser and Ellsworth, 2000) it is useful to include all data with  $CC > 0.7$  even if only recorded at one station to provide a better constraint for the whole system of equations. Figure 3a represents the distribution of correlation data graphically for event pairs with  $CC > 0.7$ . More compact clusters have event pairs that meet this criterion at more stations as expected; for example six or more stations plotted in magenta. Such a plot is useful for determining how well certain event clusters are constrained by correlation data.

**Table 1:  $CC > 0.7$  for 884 events at PANDA network\***

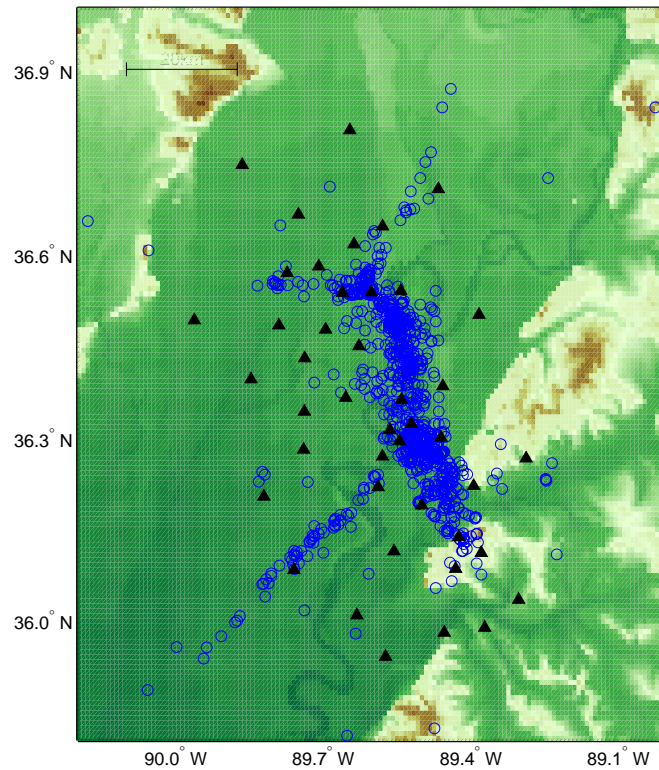
Stations	1 or more	4 or more
Number of events	830 (94%)	597 (68%)
Number of pairs	25,030	2,285
Number of observations	47,459	10,908

\*statistics for vertical component P-waves

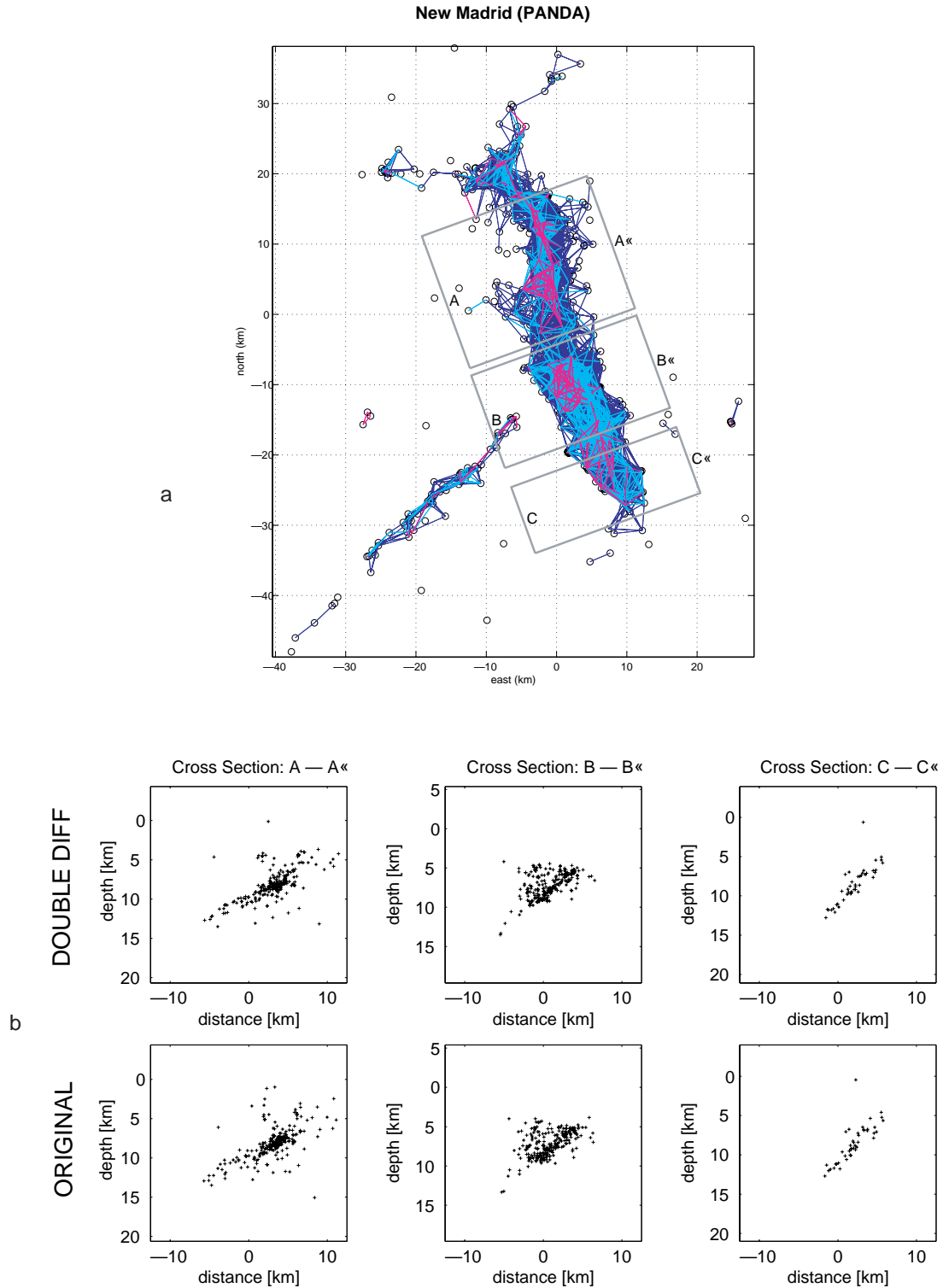
Before proceeding with the relocation stage, it is prudent to verify that the measurements are indeed improving alignments. Based on high correlation coefficients we identify a large cluster of 26 events with similar P-wave trains (Figure 4). The signal is relatively monochromatic in nature with a period of approximately 0.1 s. Figure 5 (bottom) shows that the cross correlation coefficients are quite high for all the event pairs in the cluster. Figure 5 (top) reveals a surprising feature that these events appear to be so compact spatially, but are in fact quite spread out. Some event separation distances are greater than 10 km and yet their P-waves have high correlation coefficients (Figure 5, bottom). This explains why the later S-wave arrivals on Figure 4 did not align when the P-waves are adjusted. The explanation for this behavior is probably due to the fact that there are no hard rock sites in the central Mississippi Valley, but all the PANDA stations are situated on soft sediments. These conditions may set up a site resonance for an incoming P-wave that may be less sensitive to differences in the initial locations of the events in terms of affecting waveform similarity. Whether these correlation measurements actually provide useful data to improve earthquake locations given an assumed velocity model for first arrivals should be explored more fully.

Preliminary double-difference relocation of the largest cluster (649 events) is shown in Figure 3b (upper panel), along three northeast trending cross sections. This cluster is selected to include pairs of events with at least 8 P- and/or S-wave cross-correlation measurements, to ensure a well-conditioned system of double-difference equations. The velocity model used for relocation is from Chiu et al. (1992). Weighted RMS residuals decrease from 230 ms before relocation to 34 ms after relocation. The average horizontal error determined by a bootstrap analysis of the final residual vector at the 90% confidence level is 90 m, the vertical error is 101 m. Relocation results clearly delineate a narrow zone of active faulting, with some diffuse seismicity occurring away from the main fault. The

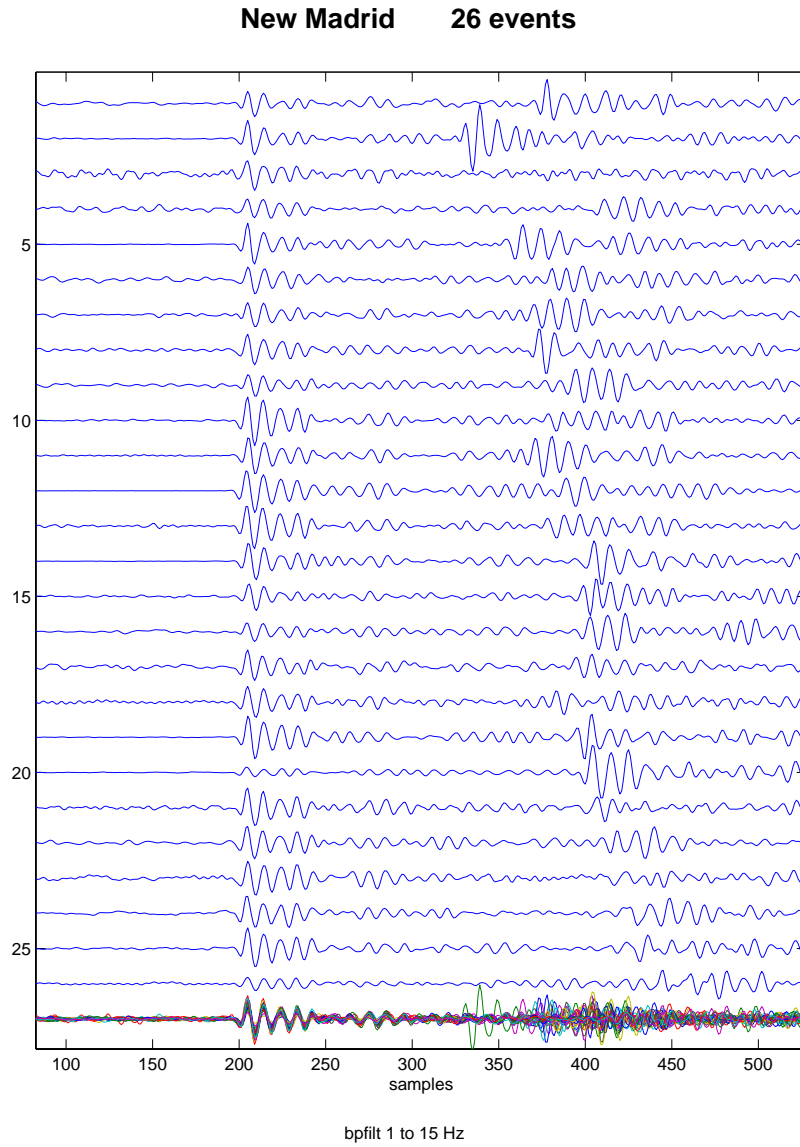
main seismically active fault plane dips to the southwest, with a dip of about  $30^\circ$  (cross section A-A') at the northern end of northwest striking step over seismicity and about  $45^\circ$  at its southern end (cross section C-C'). Comparison of the double-difference results with corresponding original network locations (lower panel in Figure 3b) indicate a sharper picture of seismicity in the new locations. Additional work is necessary to ensure proper modeling of the complicated phases (Chiu et al., 1992), such as the use of 3D models to predict differential times and partial derivatives for each event.



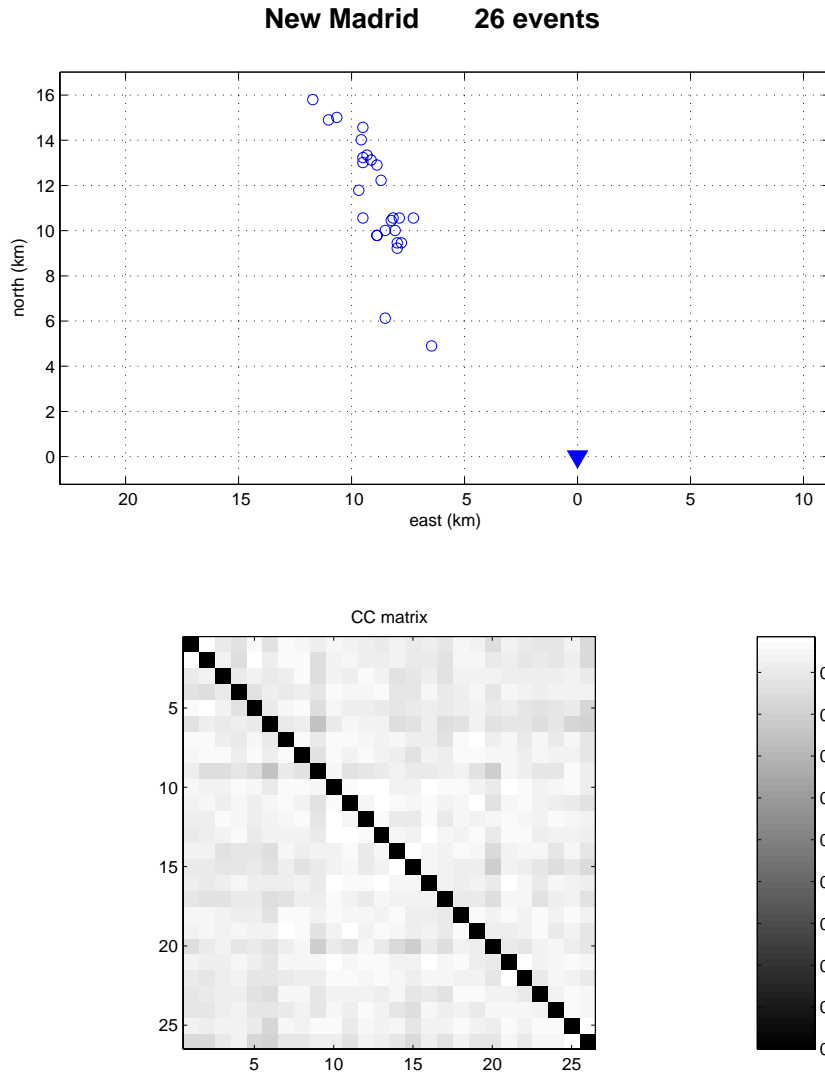
**Figure 2. 918 earthquakes (circles) that occurred during 1989-1992 in the New Madrid seismic zone recorded by the PANDA network (triangles).**



**Figure 3. a) Connectivity of correlation measurements that have  $CC > 0.7$  for event pairs at 2 (blue) or more stations, 4 (cyan), and 6 (magenta). Event locations are from the initial catalog. Boxes include events shown in (b). b) Cross sections of double-difference relocation of 649 events in the step-over zone using the cross-correlation based differential times (top row). Original locations (bottom row) are shown for comparison. Locations of cross-sections are indicated in (a).**



**Figure 4. Seismograms for 26 events at the PANDA station, i01, band-pass filtered from 1 to 15 Hz and aligned by cross correlation on the P-wave train coming in at 200 samples ( $\Delta = 0.01$  s).**



**Figure 5. (top) Map view of the epicenters for the 26 event cluster along with station i01 plotted at the origin (triangle). (bottom) Cross correlation coefficient matrix for the waveforms in Figure 4.**

#### *Charlevoix seismic zone*

Preliminary results from cross-correlating events in the Charlevoix seismic zone (Figure 6) are summarized in Table 2. Phase data was available for 2797 events. Waveforms were available for 2470 of those events. In this case, only 255 or (10%) of the events meet the criterion of  $CC > 0.7$  at four or more stations. The primary reason for this is that only 8 stations are close to the majority of the seismicity as compared to the high station density of the PANDA network (see Figure 2). This affects also the number of pairs and the number of observations. To compare networks with variable station density it may be more appropriate to consider the “1 or more stations” value which is 78% for events in Charlevoix and 94% for the PANDA network (Tables 1 & 2). Figure 6 displays the connectivity plot for the correlation data and shows hardly any event pairs with 6 or more stations recording. Another reason for the relatively low “4 or more” 10% of the events that are expected to have locations improved by correlation measurements, is earthquake density. If event-pair separation distances are too great, the waveform similarity may breakdown to the point of being unusable for accurate differential time measurement. For comparison 92% of the

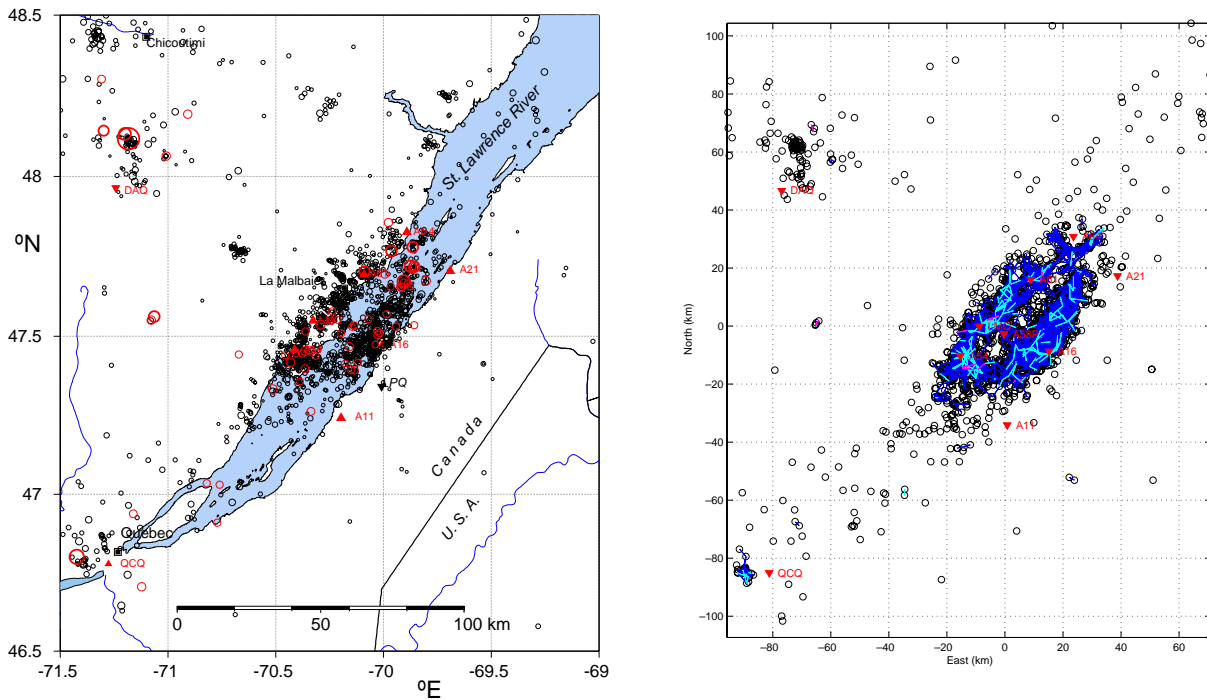


8000 events on the Calaveras fault in California were able to have significantly improved locations with the inclusion of correlation data (Schaff et al. 2002). These events were spread over a 35 km stretch of fault of minimal width (~75 m). In the Charlevoix seismic zone there are only 2470 events, which are distributed in two wide bands spanning 70 km in length and about 40 km in width. Estimating the difference in earthquake density for the two regions using these rough numbers is well over a factor of a thousand. Although this analysis is imperfect since denser clusters are likely to exist for both regions, it still gives an idea of overall improvement for the majority of the seismicity on a larger scale. In summary both earthquake density and station density are important for applying waveform cross correlation. A final comparison can be made with all of northern California where both earthquake density and station density are high. A recently completed project performing 26 billion correlation measurements for 225,000 events demonstrated that a remarkable 90% of the events in northern California have  $CC > 0.7$  with at least one other event at four or more stations (Schaff and Waldhauser, in prep.).

**Table 2:  $CC > 0.7$  for 2470 events at Charlevoix\***

Stations	1 or more	4 or more
Number of events	1916 (78%)	255 (10%)
Number of pairs	17,477	256
Number of observations	22,437	1,123

\* statistics for vertical component P-waves



**Figure 6. (left) Charlevoix seismicity and station locations. (right) Connectivity of correlation measurements that have  $CC > 0.7$  for event pairs at 2 (blue) or more stations, 4 (cyan), and 6 (magenta). Event locations (2797) are from the initial catalog for the years 1988 to 2003. Stations plotted are those that had waveforms with correlation measurements.**

## **CONCLUSIONS AND RECOMMENDATIONS**

Cross-correlation measurements for earthquakes in the New Madrid seismic zone (NMSZ) and the Charlevoix seismic zones (CSZ) indicate that a much higher number of events (68%) correlate in the NMSZ, compared to only 10% of the events that correlate in the CSZ. The reason for this discrepancy may be due to: differences in the network geometry and number of stations; differences in the type and diversity of faulting associated with the events; the variation of geophysical properties in general; and the degree of structural heterogeneity in particular, within the areas of investigation; or a combination of all. Our observations have the potential to address important scientific questions, as well as technical questions related to the earthquake location problem. We will focus on these issues during the second half of our project.

We recommend that for regions of high seismicity within which a high percentage of events cross-correlate at enough stations to achieve precise relocations, consideration be given to a wholly different paradigm for event location — namely, a framework in which events are located using cross-correlation measurements obtained from relevant portions of the waveform, rather than using phase picks.

## **ACKNOWLEDGEMENTS**

Network operators who agree to provide us with all their data are engaging in a highly civilized activity. The data cost millions of dollars to acquire and many man-years of work, and have been given us at no charge. It is typically a significant effort to provide the data in formats we can use. We greatly appreciate the assistance of the Geological Survey of Canada, and personnel at St. Louis University and Memphis State University. In particular we thank Drs. David McCormack and Jim Lyons at the Geological Survey of Canada, Ottawa, who helped assemble phase and waveform data from earthquakes in the Charlevoix region. We thank Drs. Jer-Ming Chiu of CERI, University of Memphis and Jiakang Xie at Lamont-Doherty who helped us to acquire phase and waveform data from the PANDA New Madrid deployment. We also thank Drs. Mitch Withers at CERI, University of Memphis and Robert Herrmann at St. Louis University for New Madrid seismic network data and Central Mississippi Valley seismic bulletin, respectively.

## **REFERENCES**

- Chiu, J.M., A.C. Johnston, and Y.T. Yang, Imaging the active faults of the central New Madrid seismic zone using PANDA array data, *Seismol. Res. Lett.*, 63, 375-393, 1992.
- Geller, R. J., and C. S. Mueller (1980). Four similar earthquakes in Central California, *Geophys. Res. Lett.*, 7, 821-824.
- Schaff, D. P., and P. G. Richards(2004a), Lg-wave cross correlation and double-difference location: application to the 1999 Xiuyan, China, sequence, *Bull. Seism. Soc. Amer.*, in press.
- Schaff, D. P., and P. G. Richards (2004b), Repeating seismic events in China, *SCIENCE*, 303, 1176D1178.
- 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 (2004). Optimizing correlation techniques for improved earthquake location, *Bull. Seismol. Soc. Amer.*, 94, 705-721.
- Schaff, D.P., and F. Waldhauser. Improved Differential Travel Time Measurements at the Northern California Seismic Network, *Bull. Seismol. Soc. Amer.* in prep.
- Waldhauser, F. and W.L. Ellsworth (2000), A double-difference earthquake location algorithm: Method and application to the northern Hayward fault, *Bull. Seism. Soc. Amer.*, 90, 1353-1368.