GLOBAL GROUND TRUTH DATA SET WITH WAVEFORM AND IMPROVED ARRIVAL DATA

István Bondár[1], Ben Kohl[1], Eric Bergman[2], Keith McLaughlin[1], Hans Israelsson[1], Yu-Long Kung[1], Paul Piraino[1], and Bob Engdahl[3]

Science Applications International Corporation[1]
Global Seismological Services[2]
University of Colorado at Boulder[3]

Sponsored by Air Force Research Laboratory

Contract No. FA8718-04-C-0020

ABSTRACT

The main objective of the three-year research project is to produce a quality controlled global GT0-5 event set, accompanied with waveform and groomed arrival time data sets. Our efforts are directed toward developing and refining methodologies for generating new GT events through multiple event location analysis.

Multiple event location techniques, such as Hypocentroidal Decomposition (HDC) (Jordan and Sverdup, 1981; Engdahl et al., 2004), provide precise relative locations within an event cluster. However, the absolute locations could still be biased. In order to get accurate absolute locations, independent GT information is needed. We have developed a novel multiple-event location technique, Reciprocal Cluster Analysis (RCA), which combines local data with regional/teleseismic HDC results and uses local stations as GT0 constraints to obtain accurate absolute locations.

We have validated the HDC-RCA methodology using an event cluster of GT0 nuclear explosions and GT5 earthquakes which occurred within the Nevada Test Site (NTS). We demonstrated that the HDC-RCA method requires neither dense local networks, nor prior GT information. It relies on a few local stations, provided that the station centroid is inside the event cluster. We showed that absolute locations obtained from the HDC-RCA analysis are consistent with the true GT locations as RCA reduces the regional/teleseismic bias to less than 5 km. Monte Carlo simulations demonstrated that the RCA error ellipses are conservative estimates of the absolute location uncertainties. This allows us to identify GT5 events based on the semi-major axis of their error ellipses scaled to the 95% confidence level. Using this criterion, we identified 21 out of 24 GT events in the NTS cluster.

The size of the 95% confidence error ellipses is mainly driven by the reading errors. We utilize waveform cross-correlation to reduce reading errors, and possibly identify phases not reported in bulletins. Waveform correlation also offers a way to flag and correct phase identification errors. We follow a rigorous statistical approach by using the significance of the cross-correlation to assess the similarity of waveforms. Arrival times are automatically adjusted according to the optimal alignment derived from the waveform cross-correlation, thus resulting in accurate phase picks with reduced measurement errors.

We further demonstrate the potential of the HDC-RCA approach on selected event clusters (Chi-Chi, Taiwan; Afar triangle, Africa) and we are prepared to process candidate event clusters selected from an updated EHB (Engdahl et al., 1998) bulletin.

OBJECTIVE

The main objective of the research project is to produce new ground truth events of GT5 or better quality from an updated EHB (Engdahl et al., 1998) on a global scale. In order to achieve this goal we develop a novel method, the HDC-RCA analysis, which will allow us to identify new ground truth events without the reliance on dense local networks and prior GT information. To facilitate the HDC-RCA analysis, we develop a statistically robust waveform correlation technique to obtain a consistent set of refined arrival times with reduced measurement errors.

RESEARCH ACCOMPLISHED

During the first year of the project our primary focus was to develop, test and validate the methodologies we will use to generate new ground truth events using an updated EHB (Engdahl et al., 1998) bulletin. These include the Reciprocal Cluster Analysis (RCA), a novel multiple-event location method that combines local data with regional/ teleseismic HDC results and uses local stations as GT0 constraints to obtain accurate absolute locations, as well as a statistically solid waveform correlation technique to obtain a consistent set of refined arrivals with reduced measurement errors.

Reciprocal Cluster Analysis

Multiple-event location techniques provide precise relative locations within a cluster, but the absolute locations could still be biased due to unmodeled velocity heterogeneities in the Earth. In order to get absolute locations, modern multiple-event location techniques utilize independent GT information, such as existing reference events and InSAR data (e.g., Bondár et al., 2004; Engdahl et al., 2004), seafloor bathymetry (Pan et al., 2002), and active fault lines (Waldhauser and Richards, 2004), to estimate the mislocation vector between the true and apparent cluster centroid. Therefore, the availability of accurate independent GT information limits the applicability of multiple event location methods.

Reciprocal cluster analysis is a multiple-event location technique that combines local data with the regional/teleseismic HDC results to obtain accurate absolute event locations using the local stations as GT0 constraints. The HDC method, our choice for regional/teleseismic multiple-event location, is described in detail in Jordan and Sverdrup (1981) and Engdahl et al. (2004). While HDC uses stations in the distance range 3-90° to satisfy the underlying assumption of repeating ray paths, RCA utilizes stations from 0-150 km from the hypocentroid, thus introducing new information. In the RCA inversion we fix the *pattern* of relative hypocenters and origin times obtained from HDC, and locate the station centroid of the local stations, using the relative event locations as fictitious stations. The mislocation vector between the true and apparent station centroid represents the regional/teleseismic bias in the HDC relative locations. The entire cluster is then shifted so that the apparent and true station centroids coincide, thus yielding absolute event locations. Note that locating the station centroid is equivalent to locating the hypocentroid of the cluster using the local stations. The rationale behind exploiting the reciprocity principle is that typically there are many more events in a cluster than stations; therefore, locating the station centroid represents an overdetermined problem, better posed for the inversion. Moreover, local stations are generally poorly constrained by the events (they typically suffer from huge azimuthal gaps), thus solving directly for the station centroid yields a robust solution.

The cartoons in Figure 1 illustrate the HDC-RCA procedure. First we perform an HDC analysis using regional and teleseismic stations to obtain precise relative locations within an event cluster, then we select a subset of well-connected events and local stations. We require that a local station recorded at least four events and each event is recorded by at least three local stations. The connectivity constraints ensure that we can safely fix the pattern of both events and stations, an essential requirement in the RCA analysis. Applying the reciprocity principle implies that due to the uncertainties in the relative event locations we don't know where exactly our fictitious stations are. To account for this extra error term we propagate the relative event location uncertainties into the RCA error budget. Consequently, readings for events with large relative uncertainties are downweighted in the RCA inversion. The RCA inversion provides an uncertainty estimate on the centroid shift, which is propagated back to the relative uncertainties to obtain absolute location uncertainties. Finally, we shift the entire HDC cluster (including events that were not used in the RCA analysis) with the mislocation vector between the true and apparent station centroid to obtain accurate absolute event locations, and scale the absolute error ellipses to the 95% confidence level. Events with semi-major axis less than 5 km are then promoted to GT5 status.

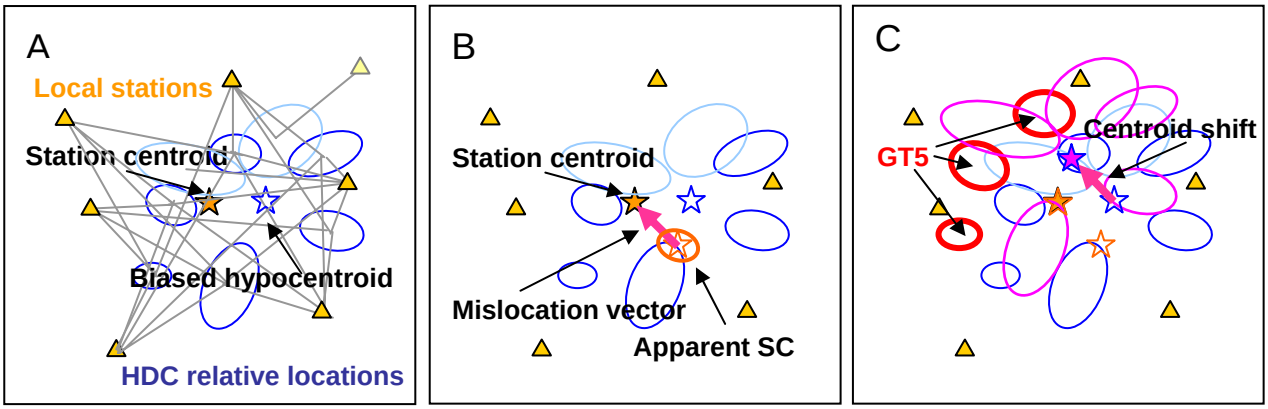

Figure 1. Illustration of the HDC-RCA analysis. A) Relative event locations and uncertainties obtained from the HDC analysis are shown in blue. Grey lines illustrate the connectivity of the cluster. Poorly connected events (light blue) and stations (light yellow) are not included in the RCA analysis. B) RCA inversion for the apparent station centroid. C) Centroid shift and GT identification.

Since RCA uses local stations to determine regional/teleseismic bias in the HDC locations, it is affected by local velocity heterogeneities. Synthetic tests showed that the local bias is bounded, as it requires 5-10% deviation from the iasp91 velocity model to build up more than 5 km bias at local distances. Using local velocity models further reduces the effect of local velocity heterogeneities.

We chose the Nevada Test Site (NTS) to validate the HDC-RCA analysis. NTS offers an ideal data set for testing and validating new algorithms as there is an abundance of GT events (both GT0 nuclear explosions and GT5 earthquakes), well-recorded at all distance ranges. Figure 2 shows our test data set selected from the SAIC ground truth database. It contains 24 GT events (10 GT0 nuclear explosions and 14 GT5 earthquakes) recorded by both local and regional/teleseismic stations. In this case we selected local stations within 50 km from the cluster centroid. Note that each event is recorded by only a subset of local stations, and while it appears a dense local network, it is also heavily unbalanced as with this local network none of the events satisfy the GT5 selection criteria of Bondár et al., (2004).

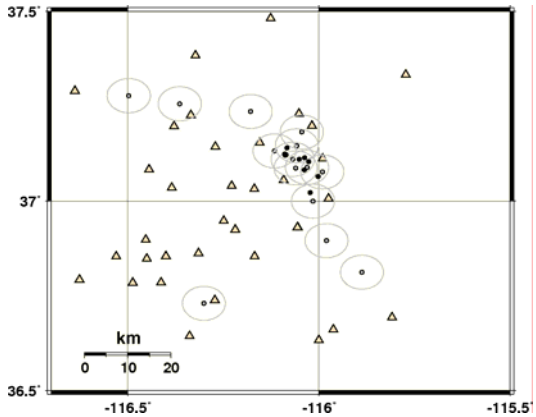

Figure 2. NTS test data set. GT0 nuclear explosions are shown as black dots, GT5 earthquakes are shown in grey, together with their 5 km uncertainty circle. Triangles indicate local stations within 50 km of the cluster centroid. Note that each event is located by a subset of local stations.

Instead of simply performing HDC on the raw data set, we first introduced an artificial bias using the Joint Hypocenter Determination (JHD, Dewey, 1972) by fixing the JHD master event to the wrong location. In this way we introduced artificial biases of 0.1° in the four cardinal directions. As a result, JHD not only perturbed the initial event locations for HDC, but more importantly, distorted the event patterns. HDC in each case removed the initial

bias, and produced nearly identical event patterns. However, the HDC hypocentroid still suffers from about 9.5 km bias. The application of RCA reduced this bias to 0.75 km. After the centroid shift, the mislocations of the GT0 events are all less than 5 km, and the absolute 95% confidence error ellipses overlap with the 5 km error circles of the GT5 earthquakes. Figure 3 shows the HDC-RCA analysis of the NTS cluster. While the locations from single-event locations using the same local network as with RCA are comparable to those obtained from the HDC-RCA analysis, the single-event location 95% confidence error ellipses are much larger, and thus none of the events could have been identified as GT5. On the other hand, with the HDC-RCA analysis we identified 21 out of 24 events as GT5. The remaining 3 events are contaminated by bad regional/teleseismic picks, resulting in large relative error ellipses.

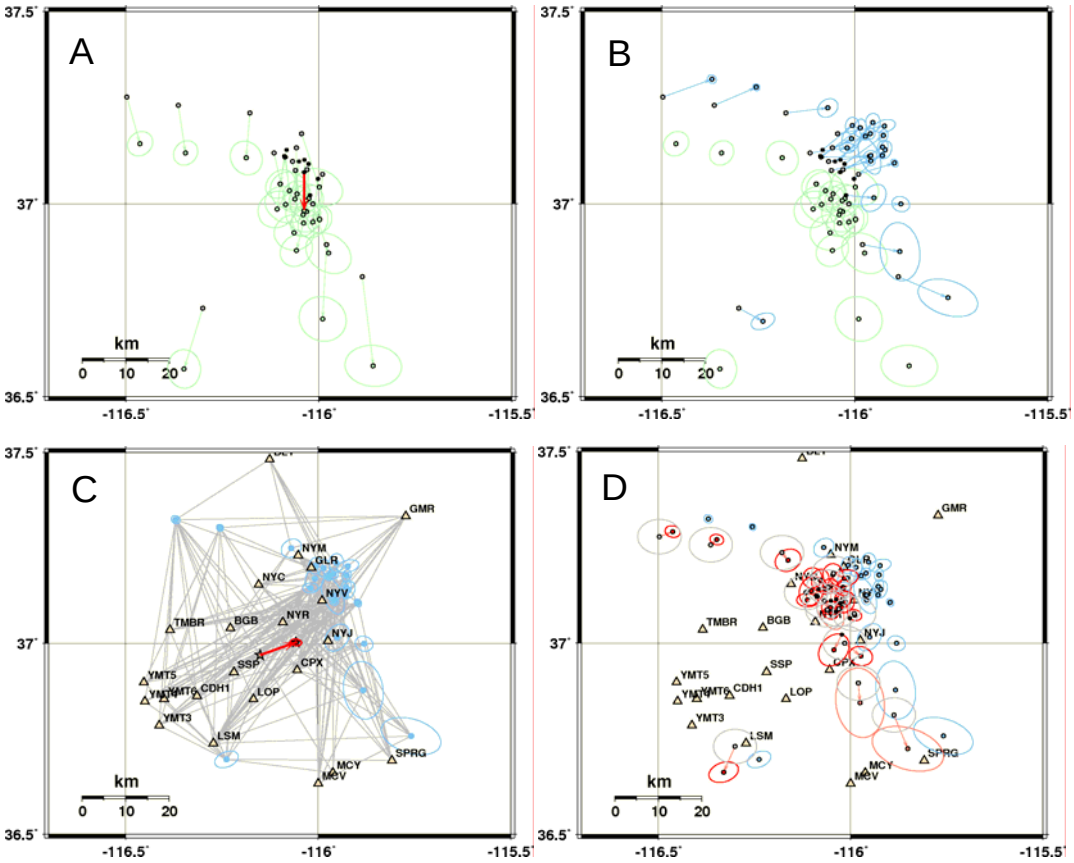

Figure 3. HDC-RCA analysis of the NTS test data set. GT0 nuclear explosions are shown as black dots, GT5 earthquakes are shown in grey. Triangles indicate local stations used in the RCA inversion. A) Initial locations (green) biased to the south by 0.1°. B) HDC relative event locations (blue). C) RCA inversion for the local station centroid. The red arrow indicates the mislocation between the true and apparent station centroid with which the entire cluster has to be shifted to obtain absolute locations. D) HDC-RCA absolute locations (red). The uncertainty in the centroid shift is propagated back to the relative uncertainties to obtain absolute location uncertainties. The absolute error ellipses are scaled to the 95% confidence level. Events with semi-major axes less than 5 km are promoted to GT5 status (bright red).

To validate the 95% confidence error ellipses obtained from the HDC-RCA analysis we performed Monte Carlo simulations using the NTS test data set. To test the sensitivity to errors in the event pattern we perturbed the relative event locations within their error ellipses and performed the RCA inversion. We also performed Monte Carlo simulations to test the effect of relative origin time and depth errors (assuming 0.2 s and 2 km depth errors, respectively). The results are shown in Figure 4. The blue error ellipse shows the uncertainty in the centroid shift scaled to the 95% confidence level, obtained from the RCA inversion using the unperturbed HDC results. The 95% error ellipse of the station centroid locations (red dots) from the Monte Carlo simulation is shown in red. The fact

that the RCA error ellipse encompasses the data cloud obtained from the Monte Carlo simulation indicates that the HDC-RCA error ellipses are conservative estimates of the absolute location uncertainties.

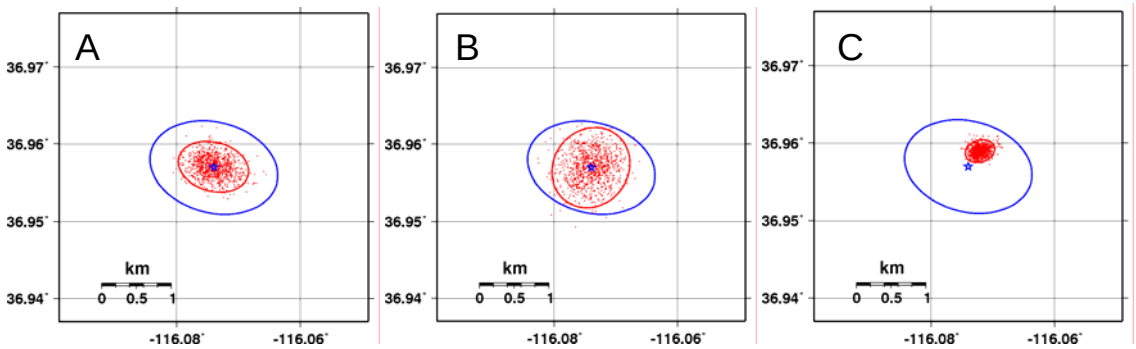

Figure 4. Monte Carlo validation of the RCA station centroid uncertainty estimate (blue). The 95% ellipses in red are derived from the data cloud (red) from the Monte Carlo simulation when A) the relative locations, B) origin times, C) depths were perturbed.

To test the robustness of results with respect to the number of local stations used in the RCA analysis we performed a bootstrap experiment on the NTS test data set. Figure 5 shows the mislocation vector between the true and apparent station centroid when only 1, 2, 3 and 4 stations are used to locate the local station centroid. Note that in the single-station case RCA degenerates to the single-event EvLoc location algorithm. The results show that once the number of local stations exceeds 3, RCA provides a robust estimate for the centroid shift. This implies that the application of RCA does not require dense local networks.

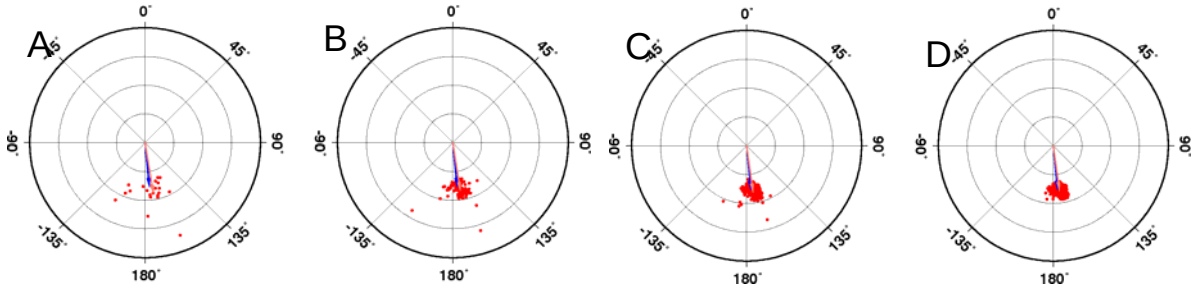

Figure 5. Station centroid mislocations (red) when only A) one, B) two, C) three, D) 4 stations are used in the RCA inversions. Concentric circles are drawn at every 10 km. The blue arrow indicates the station centroid mislocation vector when all stations were used.

We further illustrate the potential of the HDC-RCA analysis on two event clusters extracted from an updated EHB (Engdahl et al., 1998). The first example is the Chi-Chi, Taiwan aftershock sequence (Figure 6). In this cluster there are 11 events that satisfy the GT5 selection criteria of Bondár et al. (2004). We have identified 9 out of the existing 11 GT5 events, and produced 16 further GT5 events out of the 45 events used in the analysis. The HDC-RCA error ellipses overlap with the 5 km circles around the existing GT5 locations. The comparison with the Taiwan Central Weather Bureau locations shows that the HDC-RCA locations are also consistent with the local network solutions based on a local velocity model.

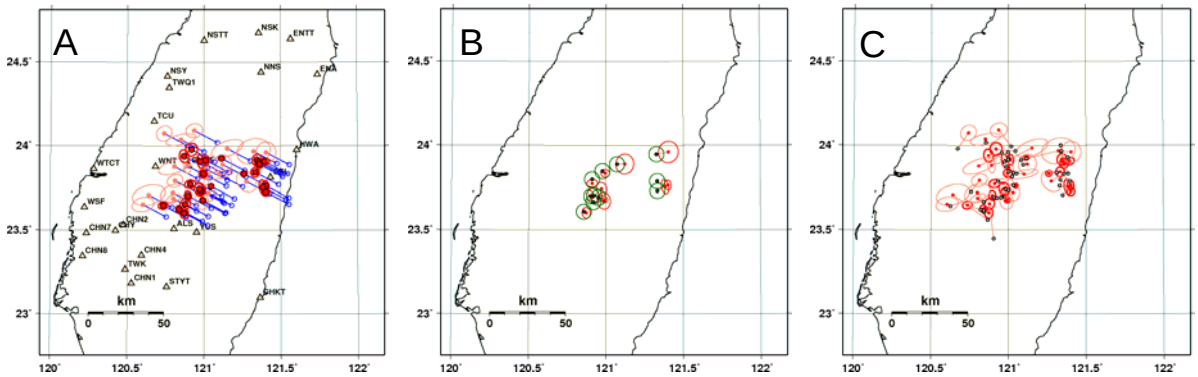

Figure 6. Chi-Chi aftershocks extracted from the EHB. A) HDC relative event locations (blue) and RCA absolute event locations (red). Events in bright red are identified as GT5. B) Comparison of prior GT5 events (green) with HDC-RCA locations (red). The HDC-RCA error ellipses overlap with the GT5 circles. C) Comparison of the Taiwan Central Weather Bureau local network locations (black) with the HDC-RCA locations (red).

Our second example is from the Afar triangle, Africa (Figure 7). In this case there is no prior GT information available. Nevertheless, with the HDC-RCA analysis we were able to identify 10 out of 18 events as GT5.

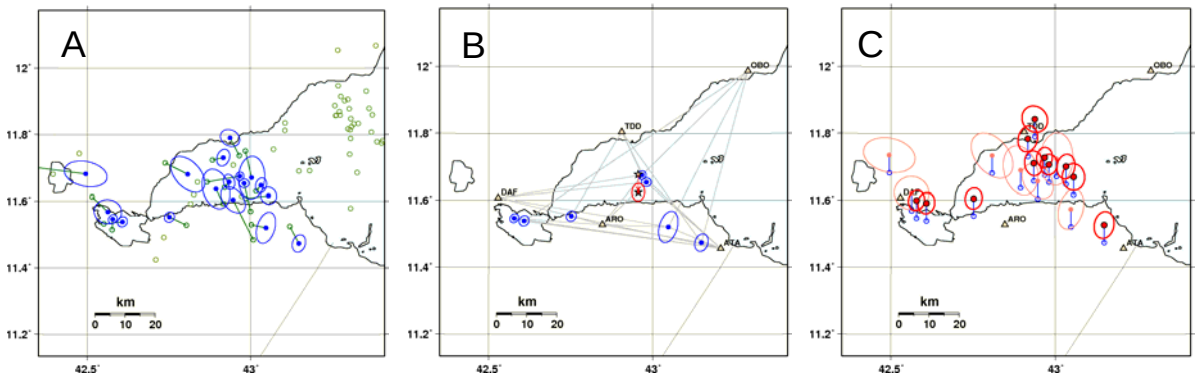

Figure 7. Djibouti cluster extracted from the EHB. a) HDC analysis. EHB locations are in green; HDC relative locations are in blue. b) RCA inversion. c) Centroid shift and GT5 (bright red) identification.

Waveform correlation

To obtain new GT5 events, the HDC-RCA methodology depends on the resulting semi-major axis of the 95% confidence ellipse being less than 5 km. The confidence ellipse is a combination of the uncertainty obtained for the station centroid in the RCA inversion and the relative location uncertainties obtained from the HDC analysis. The latter uncertainties are driven by measurement errors of regional and teleseismic phase arrivals. We employ waveform cross-correlation, which is now a proven technique for improving event locations at the local scale (Schaff et al., 2004, Shearer 1997, Thurber et al., 2003), to reduce the errors in regional and teleseismic arrival time measurements. These can in turn be fed back into the HDC-RCA analysis to obtain improved overall results.

To support the development of the methodology and to demonstrate the potential of waveform cross-correlation results to contribute to improved HDC-RCA analysis, we constructed a test dataset from events in the vicinity of the Nevada Test Site (NTS). We acquired waveform data from 595 events and 52 stations at regional and teleseismic distances with 2828 phase picks. Waveform data were acquired from the IRIS DMC and from the SMDC's archive of IMS stations and other arrays from the early 1990s through 2003 (Woodward et al., 2004).

One of the main issues we faced in this application was the very heterogeneous nature of the data set. It included broad-band three-component stations and short-period regional and teleseismic arrays with picks of regional P, teleseismic P and various secondary phases. To fully utilize these data we made three extensions to the cross-

correlation processing. We used correlation window lengths and filter bands that were phase and distance dependent. We performed the correlations on the individual channels of arrays and stations and performed a stack to obtain an array-based correlation trace. We made empirical measurement of the time-bandwidth product and used these measures to compute the statistical significance of the maximum of the correlation trace.

This last approach allowed us to replace the heuristic approach of using only those correlation results with a maximum above a given correlation threshold, typically 0.6 or 0.7. Consider that when cross-correlating noise or independent signals, Fisher's z-transform, defined as $z=\text{atanh}(r)$, has an asymptotic normal distribution with zero mean and variance, $\sigma^2=1/(N-3)$, where r is the correlation. N is the number of independent degrees of freedom which, in turn, depends on the time-bandwidth product (TB) and number of channels in the correlation stack, $N=2TB * N_{chan}$. The maximum of the correlation then follows an extreme value distribution and we compute the significance of the measured maximum from the cumulative of the extreme value distribution:

$$S = \left[\left(1+\text{erf}\left(\max(z)/\sqrt{2}\sigma_{eff}\right)\right)/2\right]^{2T_{window}\left(f_{high}-f_{low}\right)}. \qquad (1)$$

Our implementation involved using a time-domain convolution to compute the cross-correlation of the signals. We used a target window that was as much as 30 seconds longer than the template window in order to obtain cross-correlations of the template signal with uncorrelated pre-signal noise and uncorrelated coda + noise. We measured the variance of the distributions of the additional correlations to obtain an effective variance (σ^2_{eff}) used in the significance calculation. Further, we compared the variances before and after stacking the array elements to compute an effective number of channels (N_{eff}), which in turn allowed us to determine an effective time-bandwidth (TB_{eff}) product from $\sigma^2_{eff}=1/(2TB_{eff} * N_{eff}-3)$.

Figure 8 illustrates the strength of the significance test for a case where a seemingly low correlation (0.15) was found to be highly significant (0.995). These cases occur when low-SNR, long duration broadband signals (large TB_{eff}) and the full utility of an array (high N_{eff}) can be realized. Differential times were obtained from the cross-correlations by picking the lag of the maximum of the correlation with significance > 0.98. For reference, when using single channel correlations ($N_{chan}=1$), a 1 second window ($T=1$), filtered 1-10 Hz ($B=1$), a correlation threshold of 0.65 is equivalent to a significance threshold of 0.98. We inverted the measured differential times into refined absolute arrival times by using the expectation maximization algorithm to iteratively minimize the function

$$L(T;\tau) = \sum_{j=1,N}\sum_{i=1,N}\left|\tau_{ij}-\left(T_i-T_j\right)\right|^p w_{ij}, \qquad (2)$$

where τ_{ij} are the measured differential times and T_i are the absolute times. We used the normalized, z-transformed maximum correlations to weight the differential time measurements, $w_{ij} = \max(z_{ij})^2 / \sigma^2_{eff}$.

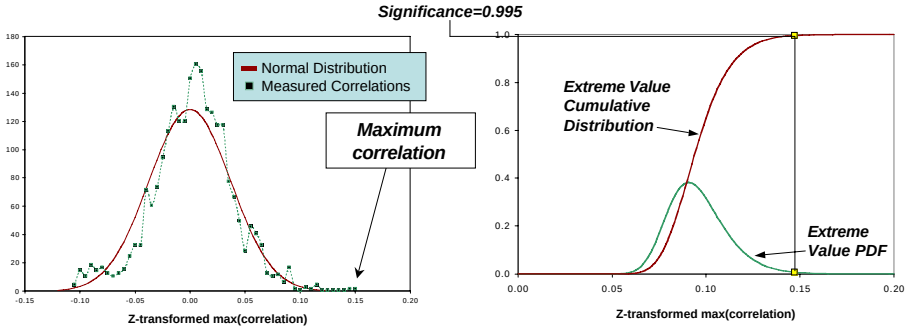

Figure 8. Example of the significance computation for a case where the maximum correlation (0.15) falls below the threshold typically used for screening correlation measurements. The significance is a measure of how much the maximum of the measured correlations (left panel, green squares) deviates from the expected maximum value (right panel, green line).

We cross-correlated all available data for the 595 events in the vicinity of NTS yielding more 400,000 correlations and differential times. After applying a significance threshold of 0.98 and inverting the differential times we obtained refined arrival times for 61% (1729 of 2828) of the arrivals. Figure 9 shows examples where the refined arrival times corrected mispicks in some cases of over 2 seconds. We found that 567 of the 595 events (95%) had at

least one arrival that was refined through our process. Of note is the fact that only about 30% of the arrivals would have been refined if we had used a threshold of 0.7 to screen out poor correlations. We plan to use the refined arrival times in the HDC analysis and evaluate the degree to which the HDC relative locations are improved by using improved arrivals.

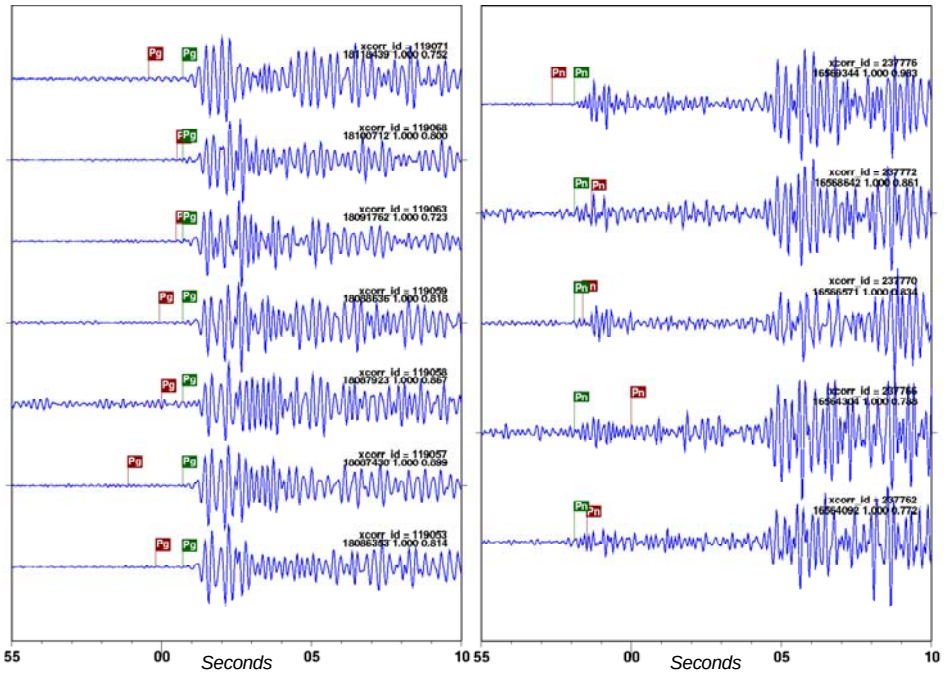

Figure 9. Arrivals and waveforms for selected phases that were refined (green flags) through the use of waveform cross-correlation. This example shows how mispicks (red flags) as large as 2 seconds were identified and corrected.

CONCLUSIONS AND RECOMMENDATIONS

We have developed a novel multiple event location technique that exploits local data to determine the regional/teleseismic bias in the HDC relative event locations. The basic idea is to fix the relative event location pattern obtained from HDC and locate the station centroid of local stations by invoking the reciprocity principle. The mislocation vector between the apparent and true (a GT0 constraint) station centroid is a robust estimate of the regional/teleseismic bias with which the entire event cluster is shifted to obtain accurate absolute locations. We have validated the method using the NTS cluster. We have demonstrated that the HDC-RCA analysis reduces the regional/teleseismic bias to less than 5 km, and that the absolute event locations are consistent with the true GT locations. We have shown that the HDC-RCA 95% confidence error ellipses are conservative estimates of the absolute location uncertainties, which allows us to promote events to GT5 status based on the semi-major axes of their 95% confidence error ellipses. The HDC-RCA analysis requires neither dense local networks nor independent GT information to produce GT5 (or better) events. Only a few stations are necessary, provided that the station centroid is inside the event cluster and has sufficient ray coverage.

We are in the process of developing applicability criteria similar to those of Bondár et al., (2004), which would guarantee that the HDC-RCA analysis produces GT5 events from a cluster. We have already identified a set of candidate metrics:

- The secondary azimuthal gap on the local station centroid defined by the events is less than a threshold;
- The local station centroid satisfies the GT5 selection criteria of Bondár et al., (2004);
- The azimuthal gap of event-station pairs, when collapsed to the origin, is less than a threshold (metrics on ray coverage)

We will perform further Monte Carlo and bootstrap experiments to refine and validate the above criteria.

We will use an updated EHB (Engdahl et al., 1998) bulletin to form event clusters for the HDC-RCA analysis. Figure 10 shows the locations of potential HDC-RCA clusters, where there are at least 5 shallow events recorded by at least 10 stations at the HDC (3-90°), and by at least 4 stations in the RCA (0-150 km) distance range.

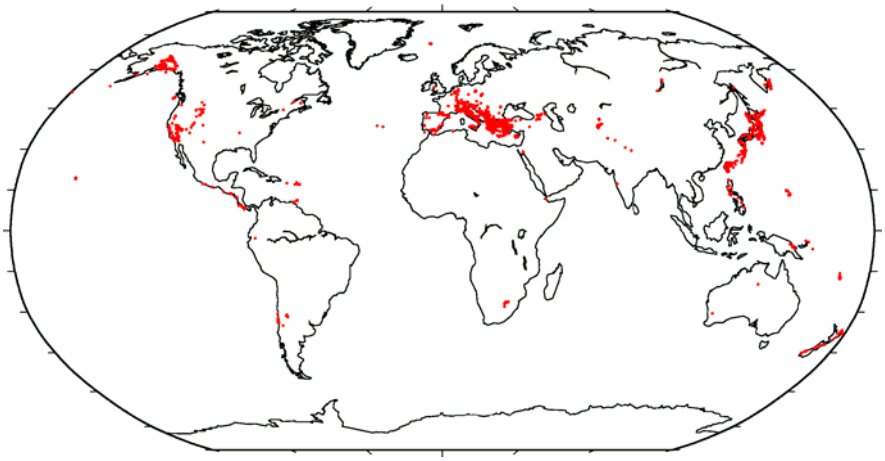

Figure 10. 1546 potential HDC-RCA clusters identified in an updated EHB bulletin.

We have developed a robust statistical framework for the waveform correlation technique, which takes advantage of arrays by stacking the correlation traces at the array elements, and more importantly, uses the significance of the maximum correlation, instead of an arbitrary threshold on the correlation coefficient, to establish a measure of similarity between waveforms. Using a significance threshold of 0.98 allows us to refine the absolute arrival times of twice as many picks as with a correlation threshold of 0.7. We will incorporate the improved waveform correlation technique in the HDC-RCA analysis.

ACKNOWLEDGEMENTS

We thank the Taiwan Central Weather Bureau for providing the local network locations of the Chi-Chi aftershock sequence.

REFERENCES

Bondár, I., S.C. Myers, E.R. Engdahl, and E.A. Bergman (2004). Epicentre accuracy based on seismic network criteria, *Geophys. J. Int.* 156: 1-14, doi: 10.1046/j.1365-246X.2004.02070.x.

Bondár, I., E.R. Engdahl, X. Yang, H.A.A. Ghalib, A. Hofstetter, V. Kirichenko, R. Wagner, I. Gupta, G. Ekström, E. Bergman, H. Israelsson, and K. McLaughlin (2004). Collection of a reference event set for regional and teleseismic location calibration, *Bull. Seism. Soc. Am.* 94: 1528-1545.

Dewey, J.W. (1972). Seismicity and tectonics of Western Venezuela, *Bull. Seism. Soc. Am.* 62: 1711-1751.

Engdahl, E.R., E.A. Bergman, S.C. Myers, and F. Ryall (2004). Improved ground truth in Southern Asia using in-country data, analyst waveform review and advanced algorithms, *Proc. 26th Seismic Research Review: Trends in Nuclear Explosion Monitoring*, LA-UR-04-5801, Orlando, FL, Sept. 21-23, 2004, Vol. 1, 257-266.

Engdahl, E.R., R.D. van der Hilst, and R.P. Buland (1998). Global teleseismic earthquake relocation with improved travel times and procedures for depth determination, *Bull. Seism. Soc. Am.* 88:722-743.

Jordan, T.H. and K.A. Sverdrup (1981). Teleseismic location techniques and their application to earthquake clusters in the south-central Pacific, *Bull. Seism. Soc. Am.* 71:1105-1130.

Pan, J., M. Antolik, and A.M. Dziewonski (2002). Locations of mid-oceanic earthquakes constrained by seafloor bathymetry, *J. Geophys. Res.* 107:B11, 2310, EPM 8 1-13, doi: 10.1029/2001JB001588.

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: Appli-cation to the Whittier Narrows, California, aftershock sequence, *J. Geophys. Res.* 102: 8269-8283.

Thurber, C. H., W-X Du, H. Zhang and W.J. Lutter (2003). Methods for improving seismic event location processing, *Proc. 25th Seismic Research Review Review - Nuclear Explosion Monitoring: Building the Knowledge Base,* LA-UR-03-6029, Vol. 1, 342-351.

Waldhauser, F. and P.G. Richards (2004). Reference events for regional seismic phases at IMS stations in China, *Bull. Seism. Soc. Am.* 94: 2265-2279.

Woodward, R., M. Bahavar and R. North (2004). Data resources to support nuclear explosion monitoring research, *Proc. 26th Seismic Research Review: Trends in Nuclear Explosion Monitoring,* LA-UR-04-5801,Orlando, FL, Sep. 21-23, 2004, Vol. 2, 790-799.