GLOBAL GROUND TRUTH DATA SET WITH WAVEFORM AND IMPROVED ARRIVAL DATA

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

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

Sponsored by Air Force Research Laboratory

Contract No. FA8718-04-C-0020

ABSTRACT

We present the final results of our three-year research project to produce a high-confidence global ground truth (GT) 5 data set. During the course of this project we have developed, tested, and validated the hybrid HDC-RCA (Hypocentroidal Decomposition and Reciprocal Cluster Analysis) methodology to produce new GT5 or better event locations from event clusters. The HDC algorithm uses regional and teleseismic data to estimate precise relative event locations with respect to the cluster centroid. The RCA algorithm uses local data to precisely locate the cluster centroid. We have demonstrated that the HDC-RCA multiple event location methodology is able to produce high-confidence GT5 (epicenter and depth) or better event locations using only a few local stations, without reliance on independent GT information. A posteriori assessment procedures and a priori applicability criteria have been developed and tested to assure the quality and high-confidence of the resulting GT5 events.

We have developed a novel, adaptive approach to waveform cross-correlation for improved differential arrival time measurements. The method finds the optimal time-bandwidth product to perform waveform cross-correlation, thus maximizing the similarity between waveforms for a wide range of seismic phases. Correlations are accepted or rejected based on their significance level derived from the estimated time-bandwidth product. We have further developed an error model to estimate the a priori uncertainties in differential time measurements in order to facilitate their inclusion with bulletin arrival time picks in the HDC algorithm. We demonstrated that differential times contribute to significant improvements in resolving the relative event locations in the HDC analysis and validated the cross-correlation differential time measurement model.

We have processed some 90 event clusters from all over the world, producing over 2,200 GT5 or better event locations at a high confidence level. The data set provides GT5 clusters in areas of the world previously devoid of GT5 reference events.

OBJECTIVE

The three year project has concluded. The objective of the research project was to produce new high-confidence ground-truth events of GT5 from an updated EHB (Engdahl et al., 1998) bulletin on a global scale. To accomplish this goal we developed a novel hybrid method, the HDC-RCA analysis, which identified new ground truth event locations without reliance on dense local networks or prior GT information.

RESEARCH ACCOMPLISHED

During the course of this project we developed, tested, and validated the hybrid HDC-RCA methodology to produce new GT5 or better locations from event clusters. We developed a novel, adaptive approach to waveform cross-correlation for improved differential arrival time measurements. Correlations are accepted or rejected based on their significance level derived from the estimated time-bandwidth product. Even a fraction of differential times can lead to significant improvements in resolving the relative event locations in the HDC analysis. Over 3,000 differential times improved the HDC event patterns of 15 clusters.

We processed 86 event clusters from around the world, producing over 2,200 GT5 or better locations at a high confidence level. The data set provides clusters of GT5 events in areas of the world previously devoid of GT5 reference events. Furthermore, by exploiting the GT5 event locations, we produced some 5,000 empirical path corrections relative to the iasp91 (Kennett and Engdahl, 1991) model, which may be used to validate 3D global velocity models.

HDC-RCA Analysis

To generate new GT5 events we developed the two-step HDC-RCA methodology. Hypocentroidal decomposition (Jordan and Sverdrup, 1981; Bergman and Engdahl, 2007) determines accurate event location patterns relative to a provisional hypocentroid using regional and teleseismic phases. Reciprocal cluster analysis (Bondár et al., 2005, 2006, 2007), using local phases only, determines the accurate location of the cluster centroid by keeping the event and station patterns fixed. Since regional and teleseismic data usually lack the resolution to resolve the full depth pattern in a cluster, event depths are typically fixed in the HDC analysis to a best educated guess, based on analysis of individual events with depth phases, waveform analyses, or prior local data. The HDC analysis produces accurate relative locations, and updates the phase identifications so that they are consistent with the fixed depth and ak135 (Kennett et al., 1995) predictions. In the RCA analysis we propagate the relative location uncertainties from the HDC results to the RCA error budget so that events with large relative errors (location and origin time) are down weighted. We assume 0.5 root mean square (RMS) second reading error for local P phases (P, Pb, Pg, Pn), and 0.9 second RMS reading error for local S (S, Sb, Sg, Sn) phases. If there are no close-in stations to the centroid, we only solve for the horizontal shift of the cluster centroid (2 unknowns) by keeping the depths and origin times fixed to the HDC results; otherwise we solve for all model parameters (horizontal, vertical, and origin time shifts, 4 unknowns). The RCA step is generally over determined (only 2 to 4 unknowns). It is imperative to use local velocity models, especially for the depth inversion. We then shift the entire cluster to eliminate the bias in the HDC cluster centroid location. Finally, in order to obtain absolute location uncertainties, the uncertainties in the hypocentroid are combined with the relative location uncertainties of individual events and scaled to the 95% confidence level.

We developed applicability criteria for RCA, which serves as an a priori test to decide if it is worth trying RCA at all. We found that the combined secondary azimuthal gap (defined as the largest secondary azimuthal gap when considering the azimuths of all event-station pairs) provides a robust metric that predicts the location accuracy of the cluster centroid. The necessary conditions below provide GT5 applicability criteria for the cluster centroid, analogous to the GT5 criteria of Bondár et al. (2004) for single-event locations.

- The combined secondary azimuthal gap is less than 140°.
- There are at least 25 station-event pairs.

The hypocentroid depth can only be resolved at a high confidence level if there are close-in stations in the cluster. If a cluster fails to satisfy the above criteria, we reject the entire cluster. We consider these criteria necessary (but not sufficient) conditions to generate GT5 events. Once the absolute location of the cluster centroid is pinned down with high accuracy, we promote events to GT5 category if the semi-major axis of their combined absolute error ellipses (HDC+RCA), scaled to the 95% confidence level, is less than 5 km. It should be noted that failing the GT5 applicability criteria does not mean that the locations are wrong; it only means that the cluster centroid cannot

be recovered with GT5 accuracy at high confidence. In many cases the cluster centroid (and thus many of the events in the cluster) could still qualify for GT10.

Since we promote locations to GT5 status based on the size of their absolute error ellipses, it is important that both the HDC relative error ellipses and the RCA absolute error ellipse on the centroid are reliable and provide the actual coverage claimed by their confidence level.

One of the fundamental assumptions in the RCA algorithm is that HDC accurately recovers the event pattern. To validate this assumption we performed HDC on the Pahute Mesa, Nevada Test Site (NTS) cluster of 52 GT0 events. HDC, using regional and teleseismic phases with ak135 (Kennett et al., 1995) travel-time tables, mislocated the event centroid by 12 km. Figure 1 illustrates that when the HDC bias is removed, the HDC event pattern (blue dots) matches the true event pattern (green dots) quite well, and the HDC relative 90% confidence error ellipses provide almost 90% coverage as expected.

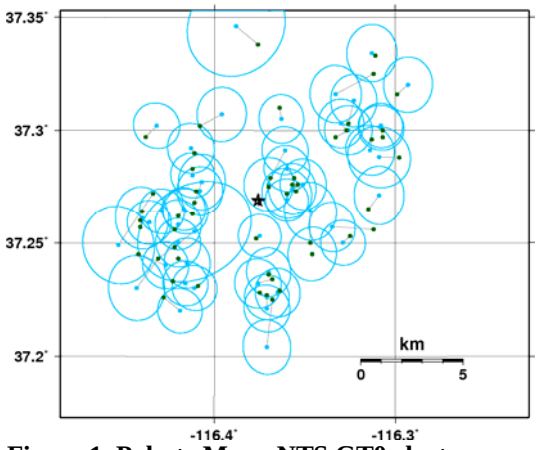

Figure 1. Pahute Mesa, NTS GT0 cluster. When the HDC location bias is removed, the HDC (blue) relative error ellipses cover 90% of the true (green) locations.

As we noted earlier, RCA eliminates the regional/teleseismic bias in the HDC hypocentroid. However, RCA is still prone to local bias due to unmodeled local velocity structure. We account for travel-time prediction errors inherent in a local velocity model by introducing generic, distance-dependent model errors, as plotted in Figure 2a. To validate that 1) the true event centroid is located within 5 km of the event centroid recovered by RCA, and 2) the RCA error ellipse covers the true centroid 95% of the time we again used the Pahute Mesa GT0 cluster. Since this time we are only interested in RCA performance and the effect of the local velocity model, we start from the true event pattern with no relative errors at all (i.e., we used the GT0 locations as the initial locations for RCA). We used the Western US velocity model by Ritsema and Lay (1995) and performed a bootstrapping RCA experiment by selecting near-optimal subnetworks with an increasing number of stations from the local network shown in Figure 2b.

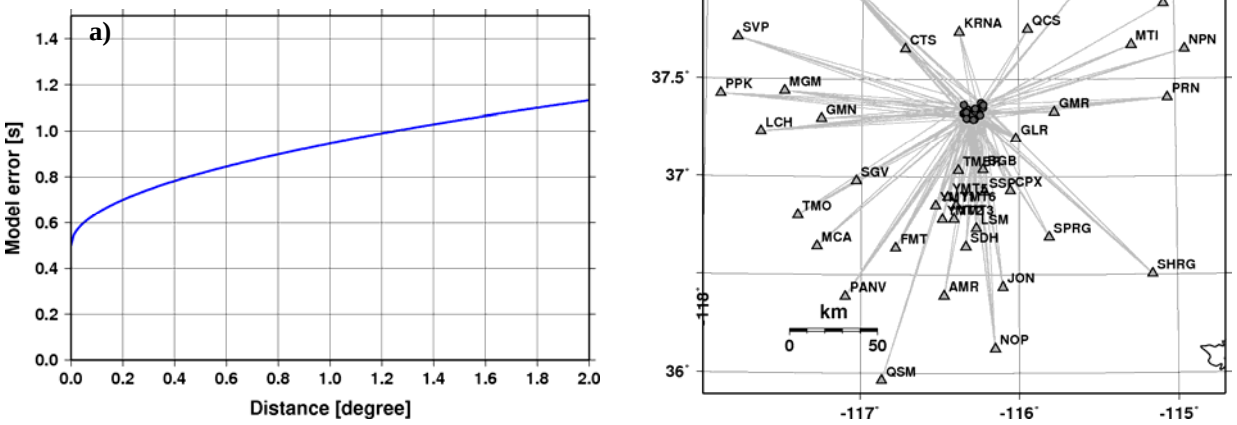

Figure 2. a) Distance dependent model error curve employed in RCA to account for local unmodeled travel-time prediction errors. b) Local network for the GT0 Pahute Mesa cluster used in subnetwork boostrap exercise.

The results are summarized in Figure 3. The local velocity model introduces a bias of about 2 km. While the event centroid mislocations are all less than 5 km (Figure 3b), the introduction of local model errors is necessary to

achieve 95% coverage. For the four cases where the centroid mislocation is larger than 5 km (orange dots in Figure 3a) the semi-major axes of the 95% error ellipses is also larger than 5 km, thus the semi-major axes of the absolute error ellipses of the individual events would also be too large to promote any events to GT5 status. Green triangles denote the bootstrap realizations for which events would be promoted to GT5 status. Blue dots represent the cases of missed GT for which the location is well within 5km, but the error ellipse is too large to be identified as GT5. This is consistent with our conservative approach to minimize the number of false alarms. In other words, we would rather loose some GT5 events than promote non-GT events to GT5 status.

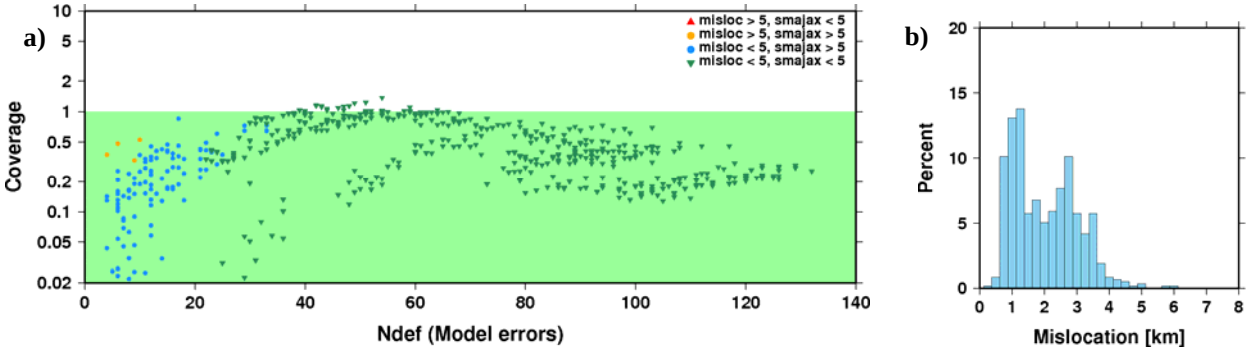

Figure 3. Results of subnetwork bootstrap exercise. a) The RCA error ellipses provide 95% coverage when local model errors are included. b) Except for 4 cases, the event centroid recovered by RCA is within 5 km of the true location.

Another assumption the RCA algorithm depends upon is that the observations are independent. With dense local networks, such as the one in Figure 2b, there is always a chance that similar ray paths produce similar travel-time prediction errors due to unaccounted local velocity heterogeneities. Thus, by ignoring the correlated error structure, the location uncertainties are underestimated. Indeed, when using the entire network (Figure 2b), the stations south of the cluster conspire to pull the RCA centroid slightly to the southeast, and the 95% absolute error ellipses (combined HDC and RCA location uncertainties) do not cover 95% of the true locations. However, when we use a subnetwork (Figure 4c) with a still acceptable combined secondary azimuthal gap, the 95% error ellipses do cover GT0 locations (Figure 4b) because the network now better satisfies the assumption of independent errors.

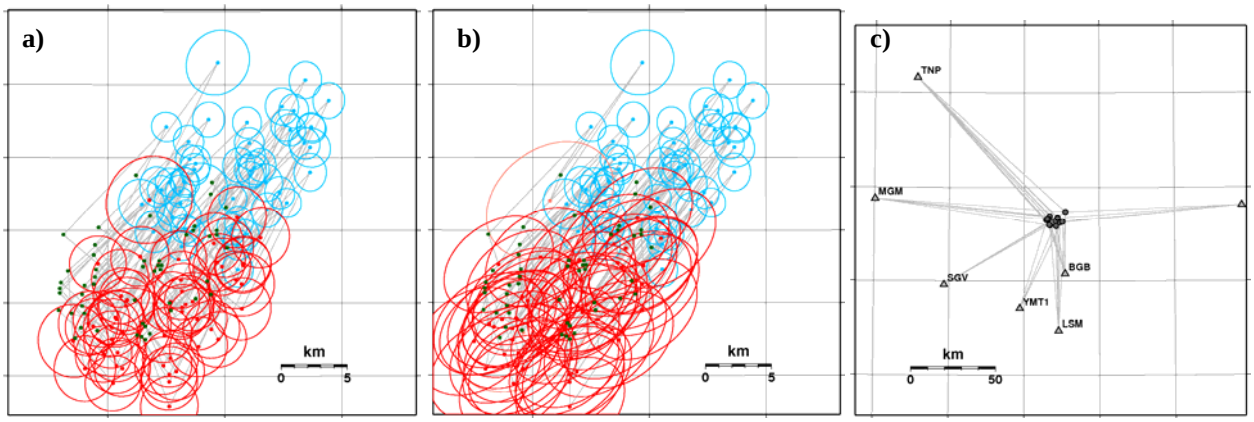

Figure 4. a) RCA results when using all stations in the local network shown in Figure 2b. Blue – HDC, red – RCA, green – GT0 locations. While the RCA event mislocations are about 3 km, the 95% error ellipses do not cover the GT0 locations. b) RCA results using the subnetwork in 4c. The error ellipses now cover the true locations, and 50 out of 52 events are identified as GT5.

Figure 4 conveniently allows us to make another important point. Since not every station recorded every event, had we had only the sparse seven-station network in Figure 4c, we could have only located four events with a conventional single event location algorithm. RCA, on the other hand, used 7 stations and 13 events to determine the HDC location bias which located all events within 5 km of GT0 and identifies 50 (out of 52) locations as GT5. Thus, in favorable conditions RCA may produce GT5 or better locations with sparse local networks.

Waveform Correlation

Significant sources of uncertainty in HDC relative locations are arrival-time measurement errors. Improvements in the HDC results may be obtained by using precise and accurate differential times from cross-correlations, either by picking the lag of the maximum from a time-domain correlation or by measuring the phase when using a cross-spectral approach (see Schaff et al., 2004 for a comparison of the techniques).

To apply the waveform cross-correlation methodology that has been very successful at local distances to regional and teleseismic data sets, we had to address several challenges. Central to exploiting the results of waveform cross-correlation is the selection of a correlation threshold. In most local applications (e.g., Schaff et al., 2004; Shearer, 1997) a heuristic approach is used to select an operational correlation threshold that was then demonstrated to be fairly robust at screening out the less reliable differential measurements. The relative uniformity of these data sets (all simple local P arrivals, typically only involving single or three-component short period data) allowed for the selection of a single uniform correlation threshold without adversely affecting the results. In our automated application, the regional and teleseismic recordings consisted of very diverse sets of stations, including single- and three-component stations as well as regional and teleseismic arrays. Data were collected from short period and broadband sensors with various responses and noise characteristics. Furthermore, regional and teleseismic recordings include a wide range of time-defining phase types (P, Pg, Pb, Pn, Sg, Sn) with varying signal bandwidths, durations and signal-to-noise ratios. Given the wide variety of time-bandwidths applicable to our data set, no single applicable correlation threshold was found.

In order to apply waveform correlation processing to a wide variety of signal bandwidths, durations, and instrument types, we developed several novel solutions. First, instead of setting an arbitrary threshold for the correlation coefficient, we measure the strength of correlation by its statistical significance. The significance of a correlation is defined as the significance level at which we can reject the hypothesis that we are correlating noise with noise in the measured time-bandwidth product (Bondár et al, 2006). This allows us to retain differential time measurements from correlation runs where the absolute correlation level might otherwise fall below the correlation threshold. In other words, a relatively low correlation may still be highly significant if the time-bandwidth product is large.

Instead of a single a priori filter and time-window we compute correlations from a suite of time-windows and a filter bank. Following Harris (1991), we measure the empirical time-bandwidth product and renormalize the maximum correlation to a value corresponding to a reference fixed time-bandwidth product. The renormalization allows us to bring all cross-correlation results into a common frame and thus set a single significance threshold. We set the reference target time-bandwidth product at 240, in which the 99.5% significance level corresponds to a correlation of 0.15. This time-bandwidth product is typical for regional arrays. Note that in a much narrower time-bandwidth of 15, which refers to parameters typically used in local cross-correlation analysis, the correlation at the same significance level would be 0.7. To optimize the significance, we compute cross-correlations over a suite of multiple filter bands and durations and then estimate the time-bandwidth for each. From the suite of cross-correlations (with different time-bandwidth products) we select the trace with the most significant renormalized correlation to measure the differential time. Alternatively, we stack the correlation suite using inverse-variance weighting, renormalize the stacked trace, calculate the significance of the maximum renormalized correlation, and if it is above the significance threshold we measure the differential time on the stacked, renormalized trace. In practice we find using the "best" correlation trace provides measurements of the differential time (less cycle skipping) as reliable as using the stack.

Figure 5 shows the distribution of "best" correlations for teleseismic P and regional Pn arrivals, i.e., those filter-band and duration combinations that yielded the highest maximum renormalized correlation above the significance threshold. As expected, the peak in the P distribution is for a short window (2.5 seconds) and the 0.8–4.5Hz filter bands. However, a large proportion (> 85%) of the correlations were optimal for a wide variety of other filter band and duration combinations, indicating that there is no single optimal filter-duration combination for all teleseismic P arrivals. The same is true for regional Pn phases.

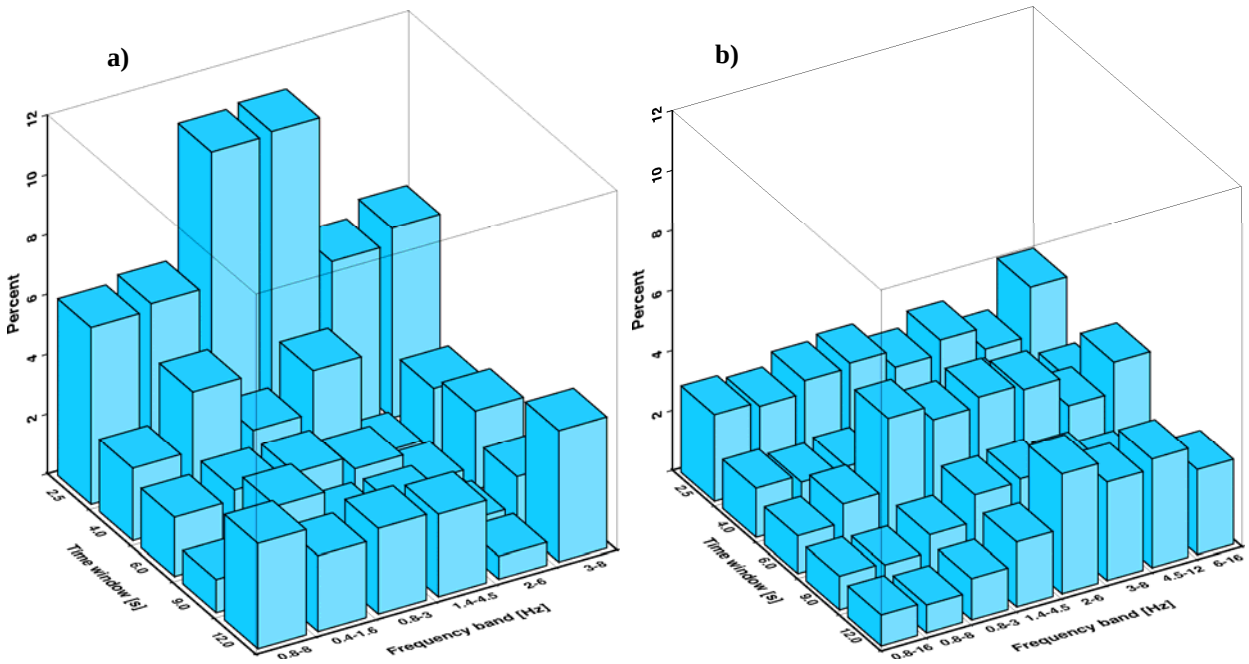

Figure 5. Histograms of best correlations in various time windows and frequency bands that yielded the highest maximum correlations above the significance threshold for teleseismic P (a) and regional Pn (b). There is no single a priori preferable filter and time window for either teleseismic P or regional Pn.

The HDC algorithm was modified to use differential time measurements along with arrival times. Because HDC now mixes the arrival and differential times, it is important to define the proper relative weighting between the arrival and differential data. This is achieved by assigning realistic measurement uncertainties to both the absolute arrival and differential times obtained from the correlation processing. Our measurement error model for differential times was empirically derived and has a power-law dependency on the renormalized Fisher Z-statistic.

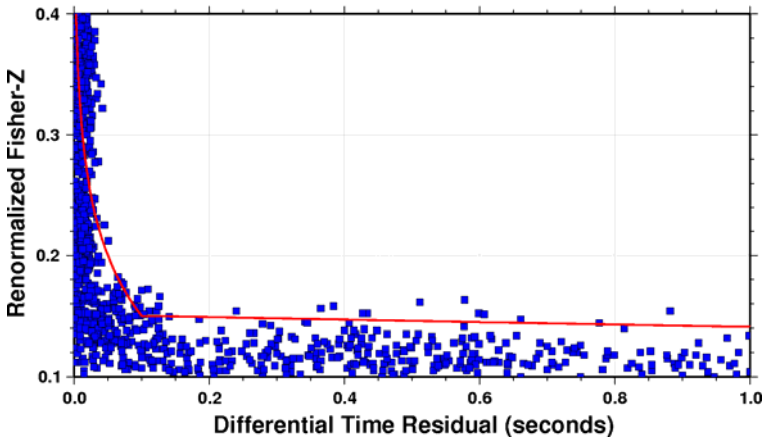

Figure 6. Differential time error model for renormalized correlations corresponding to the time-bandwidth product of 240 (typical for regional arrays). Data were re-sampled to 200 Hz prior to cross-correlation. Correlations below 0.15 (a significance level of 99.5%) are rejected.

A major challenge to quantitatively assessing a measurement error model is the lack of very accurate ground truth information, particularly for earthquake clusters, which form the bulk of the data considered in this project. To address this, we conducted a reciprocal experiment where we used clusters of seismic events (Kola Peninsula, Lop Nor and two Asian earthquake clusters) recorded at regional and teleseismic distances at arrays. We treated each cluster of events as a pseudo-array and treated the array elements as GT0 events. Figure 6 shows the distribution of maximum correlations (presented as the renormalized Fisher-Z value, corresponding to a time-bandwidth product of 240) as a function of the differential time residual. The curve in red is our derived measurement error model for $Z \geq 0.15$:

$\sigma_{\delta\tau} = \max\left(0.1 * 2^{-(z-0.15)/0.05}, 0.005\right)$. Note that differential time errors cannot be smaller than the sampling interval (as we perform cross-correlation processing in the time domain, we resample the waveforms at 200 Hz).

Figure 7 demonstrates the utility of using differential times in HDC processing. The figure shows the HDC locations for the Scotty's Junction, Nevada cluster without (Figure 7a) and with (Figure 7b) 396 differential P, Pn and Sn times. Note that the number of differential times is small compared to the 4199 bulletin picks. Even though less than 10% of the data are differential time measurements, due to their preferential weighting they yield tighter clustering of event locations and sizeable reductions in relative location uncertainties.

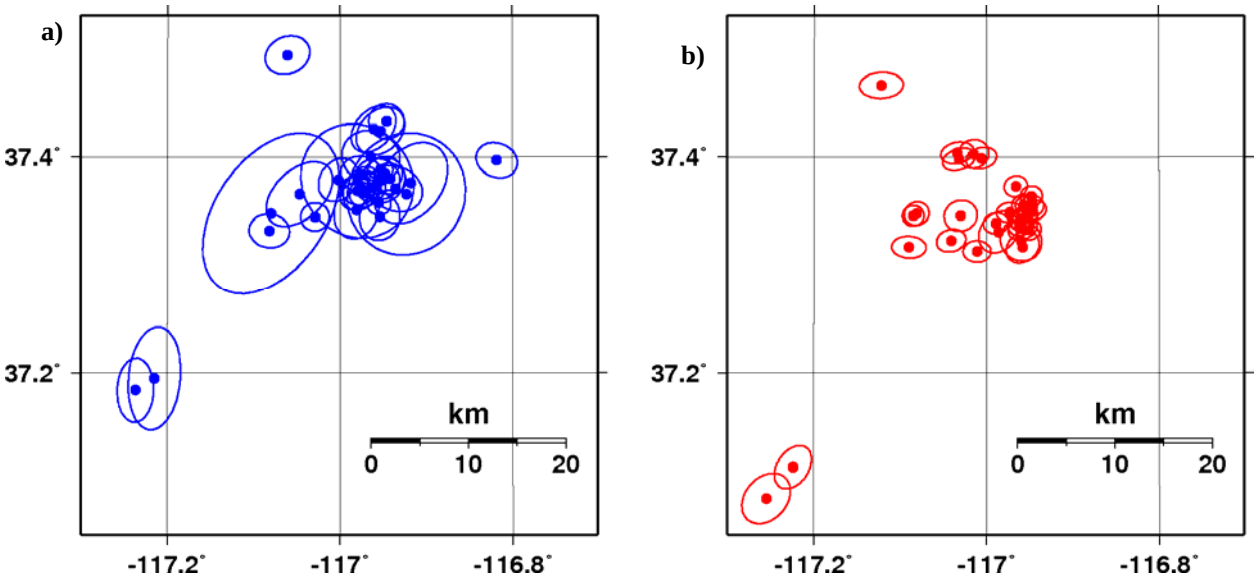

Figure 7. HDC locations for Scotty's Junction, Nevada, cluster when using a) bulletin picks only and b) both bulletin data and differential times. Even a small proportion of differential times improves the event pattern and reduces relative location uncertainties.

Figure 8 shows the distribution of *a posteriori* differential time residuals for both the bulletin picks (blue) and differential times obtained from cross-correlation processing (red). When cross-correlation differential times are used in the HDC analysis (Figure 8b), HDC obtains a much tighter fit to the cross-correlation differential times, without distorting the distribution of bulletin differential residuals.

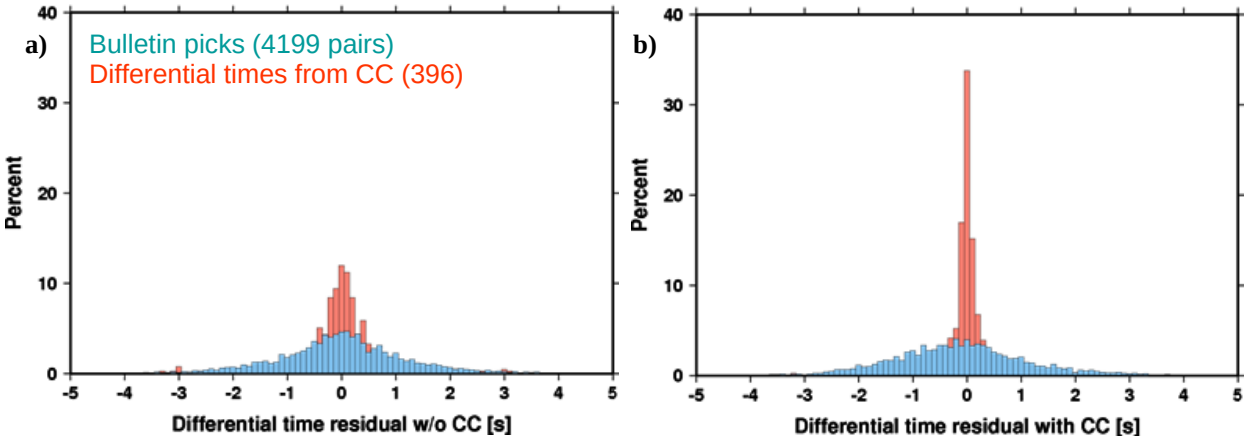

Figure 8. *A posteriori* distributions of differential time residuals of bulletin picks (blue) and those from waveform correlation (red) when using a) bulletin picks only and b) both bulletin data and differential times from cross-correlation analysis. The distribution of differential residuals derived from bulletin picks is nearly identical for both cases.

HDC-RCA Example

The Rogun, Tajikistan, cluster is located between the South Tien Shan and the Northern Pamir mountains. This region is characterized by the shallow seismicity along the Vaksh and Darvaz faults (Pegler and Das, 1998). The

HDC cluster (Figure 9a) consisted of 24 events with 376 regional/teleseismic stations. Altogether 13 events recorded by only 3 stations passed the RCA connectivity tests (Figure 9b). Even though we had only 3 stations, the geometry is so favorable that the combined secondary azimuthal gap was 110°. We used a local velocity model by Hamburger et al. (1993) to predict travel-times for the 44 Pg and Sg readings. Since we had no close-in stations, we ran RCA with a fixed depth solution (hypocentroid depth at 12 km). RCA shifted the entire cluster to the Vaksh river valley (Figure 7c) which is the surface expression of the Vaksh fault. The HDC-RCA analysis resulted in 17 GT5 events.

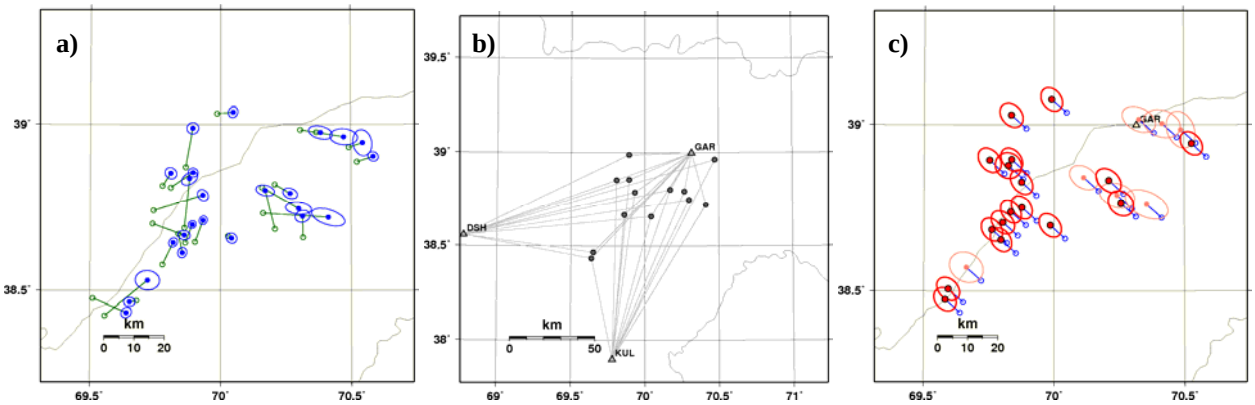

Figure 9. Rogun cluster, Pamir Mountains. a) HDC (blue) improves the event pattern with respect to the EHB single-event (green) locations. b) RCA geometry. c) RCA aligns the cluster with the Vaksh river valley.

The real strength of the HDC-RCA methodology is manifested when only a few local stations are available. Richards et al. (2006) point out that in order to achieve improved locations for more than 25% of the events in a cluster using the double-difference algorithm with differential times from waveform cross-correlation, a high local station density (one station per 100 km^2 and about 12-km distance between stations) is required. Because of the natural separation of tasks in the HDC-RCA methodology (HDC resolves the event pattern using regional and teleseismic stations, RCA reduces the bias in the hypocentroid using local data), the applicability of RCA is not restricted by such strong conditions on station density. As long as the RCA geometry is favorable, HDC-RCA is capable of producing GT events even with very sparse local station coverage.

CONCLUSIONS AND RECOMMENDATIONS

During the course of the project, we have developed a novel multiple event location method, the hybrid HDC-RCA algorithm. We have shown that HDC-RCA neither relies upon the existence of dense local networks, nor upon the existence of prior GT information. GT5 events identified by HDC-RCA are consistent with previously determined GT information. The methodology is capable of producing GT5 or better events from event clusters where other methods would not.

We applied our cross-correlation methodology to all clusters for which waveform data could be obtained at regional and teleseismic distances. We collected waveforms from both International Monitoring Station network stations and from the Incorporated Research Institutions for Seismology waveform repository for as many event-station pairs as possible. Altogether we processed about 870,000 pairs of arrivals from 47 clusters yielding about 9,200 significant correlations. We used the differential time measurements in HDC analysis for 15 clusters. For these 15 clusters, the large scale correlation processing produced 4,709 significant correlations, and thus accurate differential time measurements, of which 3,624 were used in the HDC analysis after outlier rejection.

We processed 86 clusters with the hybrid HDC-RCA analysis. Most of the clusters were extracted from an updated EHB (Engdahl et al., 1998) bulletin while in a few cases we added local data (e.g., aftershock deployments) not reported to the International Seismological Centre. We primarily focused on areas with sparse local networks. These are regions where the HDC-RCA methodology shows its real strength compared to other multiple event location methods. Our objective was to achieve a balanced global coverage of GT5 events. We also re-analyzed several classic clusters for additional cross-validation with past work (e.g., Racha). Figure 10 shows the geographic distribution of event clusters. From the 86 event clusters, 66 clusters produced altogether 2,279 GT5 or better events. We delivered the complete catalog (CSS tables) of over 3,000 event HDC and RCA locations, phase arrivals, selected path corrections, and waveforms.

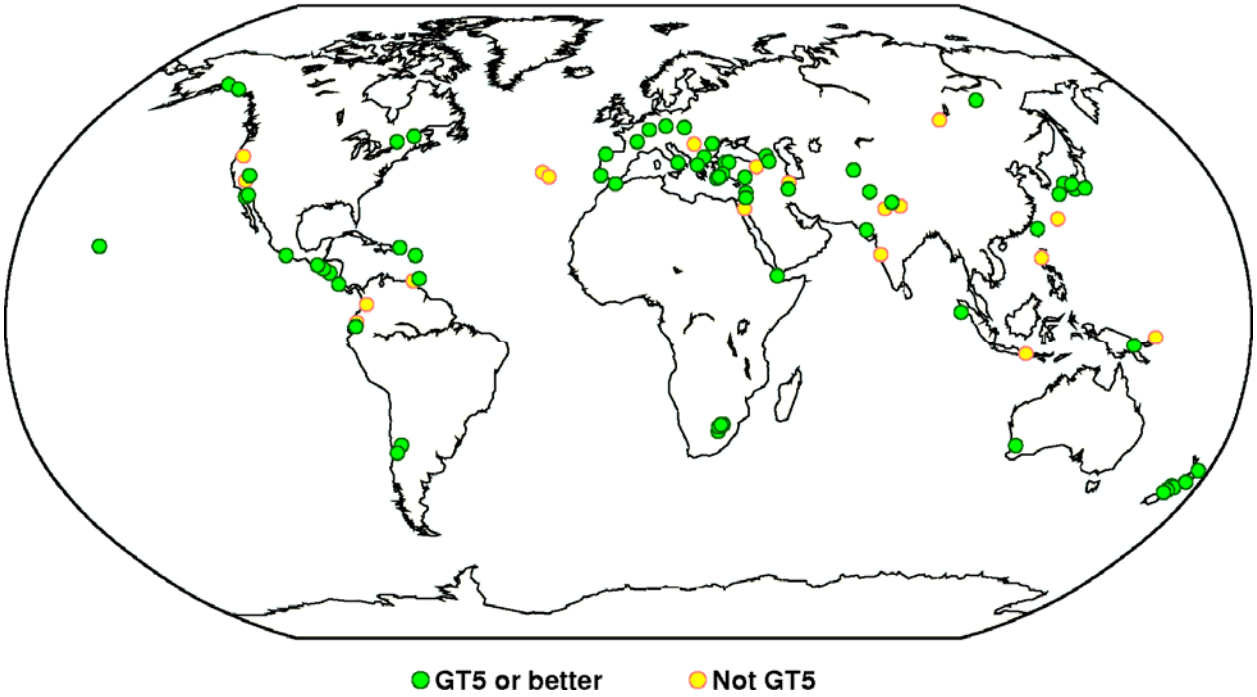

Figure 10. 86 event clusters processed during the course of the HDC-RCA GT5 project. 66 green dots represent clusters that produced GT5 or better event locations, yellow dots denote clusters that either failed the RCA applicability criteria or the 95% error ellipses were too large to promote any event locations to GT5.

REFERENCES

Bergman, E.A. and E.R. Engdahl (2007). Multiple event relocation for reference seismic events, in preparation.

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., B. Kohl, E. Bergman, K. McLaughlin, H. Israelsson, Y-L. Kung, P. Piraino, and E.R. Engdahl (2005). Global ground truth data set with waveform and improved arrival data, in *Proceedings of the 27th Seismic Research Review: Trends in Nuclear Explosion Monitoring*, LA-UR-05-6407, Vol. 1, pp.289–298.

Bondár, I., B. Kohl, E. Bergman, K. McLaughlin, H. Israelsson, Y-L. Kung, J. Given, and E. R. Engdahl (2006). Global ground truth data set with waveform and improved arrival data, in *Proceedings of the 28th Seismic Research Review: Ground-Base Nuclear Explosion Monitoring Technologies*, LA-UR-06-5471, Vol. 1, pp. 359–367.

Bondár, I., E. Bergman, E.R. Engdahl, B. Kohl, Y-L. Kung, and K. McLaughlin, (2007). A hybrid multiple event location technique to obtain ground truth event locations, submitted to *Geophys. J. Int*.

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.

Hamburger, M. W., W.A. Swanson II, and G.A. Popandopulo (1993). Velocity structure and seismicity of the Garm region, Central Asia, *Geophys. J. Int.* 115: 497–511.

Harris, D.B. (1991). A waveform correlation method for identifying quarry explosions, *Bull. Seism. Soc. Am.* 81: 2395–2418.

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.

Kennett, B. L. N. and E. R. Engdahl (1991). Travel times for global earthquake location and phase identification, *Geophys. J. Int*., 105, 429–465.

Kennett, B. L. N., E. R. Engdahl and R. P. Buland (1995). Constraints on seismic velocities in the Earth from travel times, *Geophys. J. Int*. 122: 108–124.

Pegler, G. and S. Das (1998). An enhanced image of the Pamir-Hindu Kush seismic zone from relocated earthquake hypocentres, *Geophys. J. Int.* 134: 573–595.

Richards, P. G., F. Waldhauser, D. Schaff, and W-Y Kim (2006). The applicability of modern methods of earthquake location, *Pure appl. geophys*. 163: 2–3, 351–372, doi: 10.1007/s00024-005-0019-5.

Ritsema, J. and T. Lay (1995). Long-period regional wave moment tensor inversion for earthquakes in the western United States, *J. Geophys. Res.* 100: 9853–9864.

Schaff, D., 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. (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.