ENHANCING SEISMIC CALIBRATION RESEARCH THROUGH SOFTWARE AUTOMATION AND SCIENTIFIC INFORMATION MANAGEMENT

Stanley D. Ruppert, Douglas A. Dodge, Michael D. Ganzberger, Teresa F. Hauk, and Eric M. Matzel

Lawrence Livermore National Laboratory

Sponsored by National Nuclear Security Administration
Office of Nonproliferation Research and Development
Office of Defense Nuclear Nonproliferation

Contract No. W-7405-ENG-48

ABSTRACT

The National Nuclear Security Administration (NNSA) Ground-Based Nuclear Explosion Monitoring Research and Engineering (GNEMRE) Program at Lawrence Livermore National Laboratory (LLNL) has made significant progress enhancing the process of deriving seismic calibrations and performing scientific integration, analysis, and information management with software automation tools. Several achievements in schema design, data visualization, synthesis, and analysis were completed this year. Our tool efforts address the problematic issues of very large datasets and varied formats encountered during seismic calibration research. As data volumes have increased, scientific information management issues such as data quality assessment, ontology mapping, and metadata collection that are essential for production and validation of derived calibrations have negatively impacted researchers' abilities to produce products. New information management and analysis tools have resulted in demonstrated gains in efficiency of producing scientific data products and improved accuracy of derived seismic calibrations.

Significant software engineering and development efforts have produced an object-oriented framework that provides database centric coordination between scientific tools, users, and data. Nearly a half billion parameters, signals, measurements, and metadata entries are all stored in a relational database accessed by an extensive object-oriented multi-technology software framework that includes elements of stored procedures, real-time transactional database triggers and constraints, as well as coupled Java and C++ software libraries to handle the information interchange and validation requirements. Significant resources were applied to schema design to enable recording of processing flow and metadata. A core capability is the ability to rapidly select and present subsets of related signals and measurements to the researchers for analysis and distillation both visually (JAVA GUI client applications) and in batch mode (instantiation of multi-threaded applications on clusters of processors). Development of efficient data exploitation methods has become increasingly important throughout academic and government seismic research communities to address multi-disciplinary large scale initiatives.

Effective frameworks must also simultaneously provide the researcher with robust measurement and analysis tools that can handle and extract groups of events effectively and isolate the researcher from the now onerous task of database management and metadata collection necessary for validation and error analysis. Sufficient information management robustness is required to avoid the loss of metadata that would lead to incorrect calibration results in addition to increasing the data management burden. Our specific automation methodology and tools improve the researchers ability to assemble quality-controlled research products for delivery into the NNSA Knowledge Base (KB). The GNEMRE Program software and scientific automation tasks also provide the robust foundation upon which synergistic and efficient development of seismic calibration research may be built.

OBJECTIVE

The NNSA GNEMRE Program has made significant progress enhancing the process of deriving seismic calibrations and performing scientific integration with automation tools. We present an overview of our software automation efforts and framework to address the problematic issues of improving the workflow and processing pipeline for seismic calibration products, including the design and use of state-of-the-art interfaces and database centric collaborative infrastructures. These tools must be robust, intuitive, and reduce errors in the research process. This scientific automation engineering and research will provide the robust hardware, software, and data infrastructure foundation for synergistic GNEMRE Program calibration efforts. The current task of constructing many seismic calibration products is labor intensive, complex, expensive and error prone. The volume of data as well as calibration research requirements has increased by several orders of magnitude over the past decade. The increase in quantity of data available for seismic research over the last two years has created new problems in seismic research; data quality issues are hard to track given the vast quantities of data, and this quality information is readily lost if not properly tracked in a manner that supports collaborative research. We have succeeded in automating many of the collection, parsing, reconciliation and extraction tasks individually. Several software automation tools have also been produced and have resulted in demonstrated gains in efficiency of producing derived scientific data products. In order to fully exploit voluminous real-time data sources and support new requirements for time-critical modeling, simulation, and analysis, continued expanded efforts to provide scalable and extensible computational framework will be required.

RESEARCH ACCOMPLISHED

The primary objective of the Scientific Automation Software Framework (SASF) efforts is to facilitate the development of information products for the GNEMRE regionalization program. The SASF provides efficient access to, and organization of, large volumes of raw and derived parameters, while also providing the framework to store, organize, integrate and disseminate derived information products for delivery into the NNSA KB.

These next generation information management and scientific automation tools are used together within specific seismic calibration processes to support production of tuning parameters for the United States Atomic Energy Detection System (USAEDS) run by the Air Force. The automation tools create synergy and synthesis between complex modeling processes and very large data sets by leveraging a scalable and extensible database centric framework. The requirements of handling large datasets in diverse formats, and facilitating interaction and data exchange between tools supporting different calibration technologies, has led to an extensive scientific automation software engineering effort to develop an object oriented database-centric framework using proven research-driven workflows and excellent graphics technologies as a unifying foundation.

The current framework supports integration, synthesis, and validation of the various different information types and formats required by each of the seismic calibration technologies. For example, the seismic location technology requires parameter data (site locations, bulletins), time-series data (waveforms), and produces parameter measurements in the form of arrivals, gridded geospatially registered correction surfaces and uncertainty surfaces. Our automation efforts have been largely focused on research support tools, RBAP (Regional Body-wave Amplitude Processor) and KBALAP (Knowledge Base Automated Location Assessment and Prioritization). Further, increased data availability and research requirements have driven the need for multiple researchers to work together on a broad area, asynchronously.

Database Centric Coordination Framework

As part of our effort to improve our efficiency we have realized the need to allow researchers to easily share their results with one another. For example, as the location group produces ground truth (GT) information, that information should become available for other researchers to use. Similarly, phase arrival picks made by any qualified user should also become immediately available for others to use. This concept extends to sharing of information about data quality. It should not be necessary for multiple researchers to have to repeatedly reject the same bad data, or worse, miss rejecting bad data. Rather, once data are rejected because of quality reasons they

should automatically be excluded from processing by all tools. We are implementing this system behavior using database tables, triggers, stored procedures and application logic. Although we are at the beginning of this implementation, we have made significant progress over the last year with several kinds of information sharing using the new database centric coordination framework. These are discussed below.

Significant software engineering and development efforts have been applied successfully to construct an object oriented database framework that provides database centric coordination between scientific tools, users, and data. A core capability this new framework provides is information exchange and management between different specific calibration technologies, and their associated automation tools, such as seismic location (e.g,. KBALAP), seismic identification (e.g., RBAP), and data acquisition and validation (e.g., KBITS). A relational database (ORACLE) provides the current framework for organizing parameters key to the calibration process from both Tier 1 (raw parameters such as waveforms, station metadata, bulletins, etc.) and Tier 2 products (e.g., derived measurements such as GT, amplitude measurements, calibration and uncertainty surfaces). Seismic calibration technologies (location, identification, etc.) are connected to parameters stored in the relational database by an extensive object-oriented multi-technology software framework that includes elements of schema design, PL/SQL, real-time transactional database triggers, and constraints, as well as coupled Java and C++ software libraries to handle the information interchange and validation requirements. This software framework provides the foundation upon which current and future seismic calibration tools may be based. Interim results and a complete set of working parameters must be available to all research teams throughout the entire processing pipeline. Finally, our development staff has continually and efficiently leveraged our java code library, achieving 45% code reuse (in lines of code) throughout several thousand java classes. Source code control is managed by CVS (source code) and ER Studio (schema designs).

Sharing of Derived Event Parameters

In order to calibrate seismic monitoring stations, the LLNL Seismic Research Database (SRDB) must incorporate and organize the following categories of primary and derived measurements, data, and metadata:

Tier 1: Contextual and raw data
- Station parameters and instrument responses
- Global and regional earthquake catalogs
- Selected calibration events
- Event waveform data
- Geologic/geophysical data sets
- Geophysical background model

Tier 2: Measurements and research results
- Phase picks
- Travel time and velocity models
- Rayleigh and Love surface wave group velocity measurements
- Phase amplitude measurements and magnitude calibrations
- Detection and discrimination parameters
- Integrated/merged GT data sets

Automating Tier 1

Corrections and parameters distilled from the calibration database provide needed contributions to the NNSA KB for the ME/NA/WE region and will improve capabilities for underground nuclear explosion monitoring. The contributions support critical functions in detection, location, feature extraction, discrimination, and analyst review. Within the major process categories (data acquisition, reconciliation and integration, calibration research, product distillation) are many labor intensive and complex steps. The previous bottleneck in the calibration process was in the reconciliation and integration step. This bottleneck became acute in 1998 and the KBITS suite of automated parsing, reconciliation, and integration tools for both waveforms and bulletins (ORLOADER, DDLOAD, UpdateMrg) were developed. The KBITS suite provided the additional capability required to integrate data from many data sources and external collaborations. Data volumes grew from the 11,400 events with 1 million

waveforms in 1998 to over half a billion raw parameters, measurements and associated 100 terabytes of continuous data today (e.g., Ruppert et al., 1999; Elliott et al., 2006).

Continuous seismic data automation

We receive enormous amounts of seismic data daily that must be properly processed. Previously, the movement and management of data were performed manually by our IT staff and were extremely time intensive and inefficient. In response, we designed and implemented a distributed (multi-machine), multi-process solution to help automate the collection, movement, cataloging, reporting, viewing and error processing of waveform segmentation data from multiple academic and government sources. The distributed processes are being written in Java, using encrypted data transfers, logging, an embedded Java relational database (Derby) for maintaining transfer metadata, and a monitoring interface for reporting and quality control. Also, the ability to easily query and view available continuous data was added to improve the efficiency of quality control and recording of metadata.

Automating Tier 2

As the data sources required for calibration have increased in number and source location, it has become clear that the manual, labor intensive process of humans transferring thousands of files and unmanageable metadata cannot keep the KBITS software fed with data to integrate, nor could the seismic researcher efficiently and consistently find, retrieve, validate, or analyze the raw parameters necessary to effectively produce seismic calibrations in an efficient manner. Significant software engineering and development efforts were applied to address this critical need to produce software aids for the seismic researcher. Thus, the main focus of our development efforts is on the development of two scientific automation tools, RBAP and KBALAP, for seismic location and seismic identification calibration tasks, respectively.

The RBAP Program

The RBAP is a station-centric Tier 2 automation tool; it is an interactive, graphical and highly specifiable software program that acts as a picker and an MDAC calculator (Elliott et al., 2006). RBAP helps to automate the process of making amplitude measurements of regional seismic phases for the purpose of calibrating seismic discriminants at each station. RBAP generates station centric raw, and Magnitude Distance Amplitude Correction (MDAC) corrected Pn, Pg, Sn and Lg amplitudes along with their associated calibration parameters (e.g. phase windows, MDAC values, reference events, etc.) in database tables. It strictly follows Working Group (WG) 2 standardized MDAC processing, and it replaces the original collection of LLNL scripts. RBAP has a number of advantages over the previous scripts. It is much faster, significantly easier to use, allows for collaboration, scales more easily to a larger number of events and permits efficient project revision and updating through the database.

RBAP projects are station centric; stations can be either single stations or arrays, where arrays focus on a reference element. Each project also specifies one or more regions, which can be simple rings or user-defined polygons; each region may be assigned its own velocity model. Once defined, concepts such as geographic regions are available to all researchers and all projects; interfaces include extensive use of modern mapping technologies and data tables. Table schema are driven by research workflows. RBAP makes use of the data type manager concept extensively, and includes separate managers for velocity models, regions and events. Events are shown color-coded on a map for ease of use. RBAP also includes a graphical phase picker that generates windows automatically for the Pn, Pg, Sn and Lg phases using times predicted by the velocity model. The picker is geared towards using signal-to-noise ratios for regional body wave amplitude measurements, and picks are automatically advanced according to applied velocity models. A new GIS and picking subsystem is now in place. The new GIS package will replace the MapObjects system currently in use.

Some key features of RBAP are listed below:

- Based on WG 2 standardized algorithm
- Fast and efficient calibration
- Project management
- Utilizes database for up-to-date results
- Batch processing
- Engenders collaboration, consistency and efficiency

Support for tiered projects—new in FY07

Users can now create a parent "calibration" project which would define the velocity model and MDAC parameters for a specific station and region. Typically this would be done using earthquake only data. Subsequent "child" projects could then be created for the same station and region, and MDAC corrected amplitudes would be calculated using the parent's calibrated parameters. This will allow us to explore different source types in understanding discrimination.

Adding MDAC and Coda Magnitude processing module to RBAP—new in FY07

MDAC calibration in RBAP has been enhanced with the addition of a DiscrimTool utility. The DiscrimTool is a Java based GUI program that allows a user to retrieve large blocks of raw and MDAC corrected amplitude measurements from the database based on a user defined discriminant
(e.g., Pg (6–8 Hz)/Lg (6–8 Hz)). Multiple data sets can be selected and plotted so that users can investigate different discriminants on the fly. The tool is currently being modified to allow an inversion for optimal discriminant combinations.

To supplement RBAP in source identification we are developing a waveform processing tool (WFT) with the capability of measuring amplitudes and coda magnitudes. WFT was originally written as part of the Hydroacoustic Blockage Assessment Tool (HABAT) code, but has been enhanced to work generally with seismic data and has since been used in combination with seismic inversion projects (Flanagan et al., 2006). We have recently added two programs to the WFT: the Amplitude Measurement Tool (AMT) calculates spectral amplitudes, while the Coda Tool calculates coda magnitudes for calibrated regions. The AMT (Figure 1) was designed to allow basic amplitude measurements and MDAC processing for flatfile data—specifically for cases when the LLNL database in inaccessible, or when investigating phases that are not normally dealt with (such as the hydroacoustic T phase). It performs all the basic RBAP amplitude calculation using the same MDAC parameter setup developed in RBAP, allows creation of DiscrimData tables and can plot the results. The Coda Tool (Figure 2) will allow an investigator with a basic background in coda theory to read either database or flat file seismic data and calculate source magnitudes given a regional calibration. Once data are read in, the user can calculate seismic data envelopes, calculate synthetics based on published theory (Mayeda et al., 2003), compute the spectral amplitudes, add site and path corrections and compute the final coda Mw.

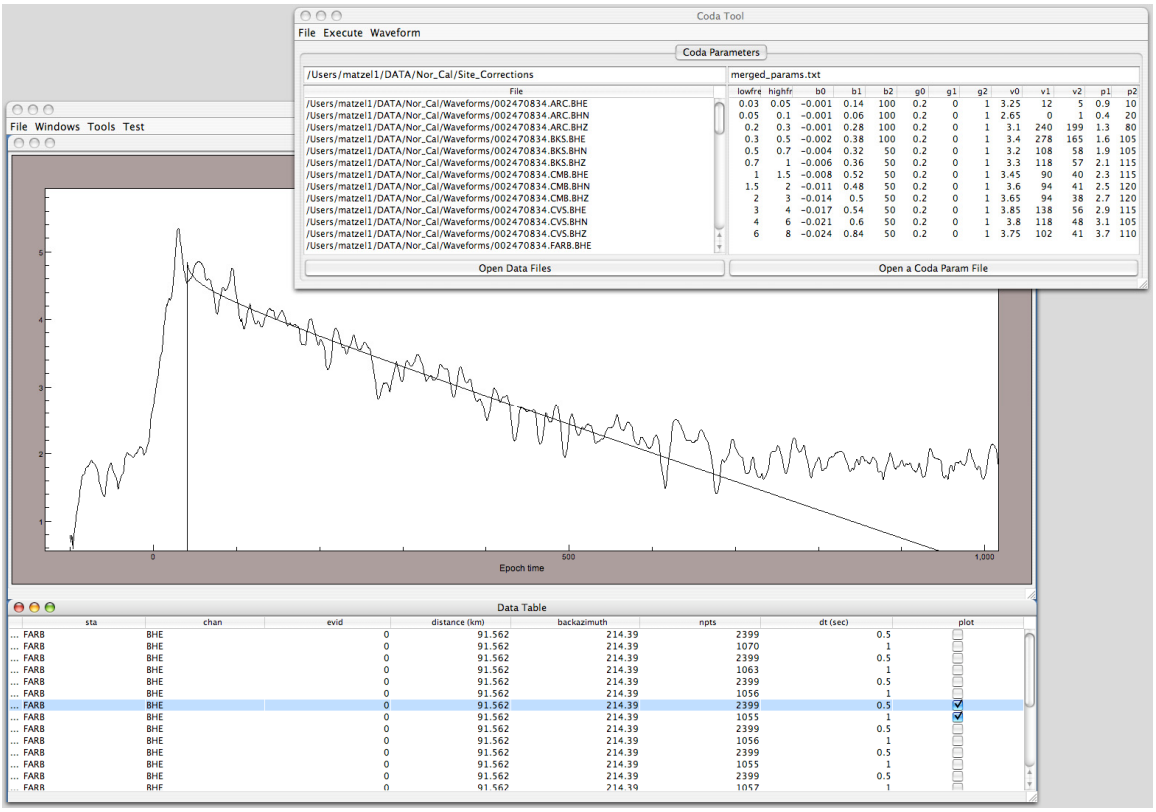

Figure 1. A screenshot of the Coda Mw measurement tool, illustrating the measured data envelopes compared with synthetic envelopes.

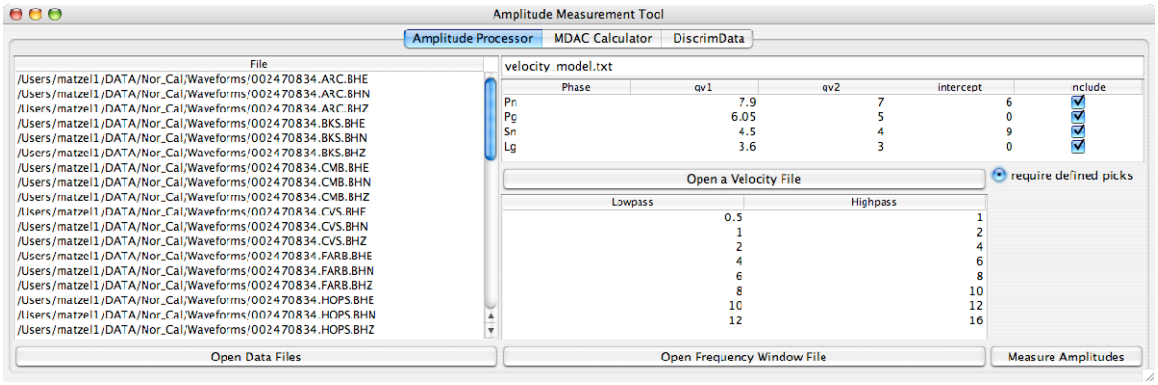

Figure 2. A screenshot of the Amplitude measurement tool, which calculates spectral amplitudes for the seismic analysis code (SAC) and CSS flatfiles and applies MDAC corrections to the results.

The KBALAP Program

The Knowledge Base Automated Location Assessment and Prioritization (KBALAP) program is another Tier 2, event centric automation effort in the GNEMRE program (Elliott et al., 2006). It is a highly interactive, graphical tool which uses a set of database services and a client application based on data selection profiles that combine to efficiently produce location ground truth data which can be used in the production of travel time correction surfaces, and as part of the preferred event parameters used by other tools in our processing framework.

KBALAP's database services are responsible for evaluating bulletin and pick information as it enters the system to identify origin solutions that meet pre-defined GT criteria with no further processing, and for identifying events that would likely meet a pre-defined GT level if a new origin solution were produced using available arrivals. The

database service is also responsible for identifying events that should have a high priority for picking based on their existing arrival distribution, and the availability of waveform data for stations at critical azimuths and distances.

The interactive portion of KBALAP has the following principal functions:

- Production of GT origins through prioritized picking and location,
- Specification of GT-levels for epicenter, depth, origin time, etype
- Batch-mode location of externally produced GT information
- Production of array azimuth-slowness calibration data
- Easy review and modification of event parameters used by all GNEM researchers

Some key KBALAP features are listed below:
- Fast and efficient location
- Project management and collaboration
- Batch processing

The Site Merge Effort—New in FY07

Information about seismic station position and installed instrumentation is to a greater or lesser extent, fundamental to all the processing done within the GNEM Program. However, despite the importance of accurate information about seismic stations, in practice it is difficult to obtain a compilation of station information that does not include errors. There are many sources for these errors, including the following:
- Imprecise surveying/reporting by station operators
- Transcription errors
- Unrecorded station movements or equipment modifications.

The situation is complicated even more by the fact that many different compilations have been produced using different sources and different assumptions, and these compilations are inconsistent with one another.

In the past, we have dealt with inconsistencies on a case by case basis. When a problem was identified, we would "fix" the offending data in our SITE table and go on. While this approach was problematic in a number of ways, given the limitations of the CSS SITE table and our need to build out other parts of our infrastructure, it was judged to be the best we could do. As the labs coordinate more in the process of producing calibration products for monitoring purposes, the need for a unified, consistent SITE table has become more apparent. Producing and maintaining such a table by integrating and reconciling our individual SITE tables is an even more difficult undertaking than simply maintaining an internal-use-only SITE table. Mainly this is because of the need to resolve conflicts in a way that is trackable, reproducible, and with documented decisions/assumptions.

We were tasked this year with performing the location ground truth merge (GT merge) between contributing laboratories. This effort depends critically on having a unified SITE table of the highest possible quality. This has accelerated our work on producing a SITE merge, and we now have a system that while still in need of further development, is adequate for purposes of the GT merge. Our merge process is implemented in Java and in PL/SQL and uses a number of tables to track metadata about the merge process. The codes allow for repeated contributions by the same author allowing, for example, updating of the merged SITE as new versions of the NEIC station book become available. The results and documentation will be provided to the relevant NNSA GNEM working groups for coordination and consideration.

Our approach to merging SITE data is to handle the position, elevation, operating epochs, station movements, array membership and possible code aliasing separately. We take this approach because there is no guarantee that a particular contributor's information about a SITE will be uniformly better or worse than information from another source.

When SITE data come into the system, they are placed into a multi-author site table (and supporting tables) that hold all the unmerged data. Before a new merge is executed, a process is run that identifies unresolved discrepancies (over a threshold value) in position and elevation. Any stations with unresolved discrepancies are added to

appropriate discrepancy tables. Although the merge can continue without resolving the discrepancies, these stations will not become part of the merged SITE table.

Discrepancies can be resolved in one of two ways; either by making entries in a preferred position (or preferred elevation) table or by making entries in a rejected position (elevation) table. The reason column in each of these tables allows up to a 2000 character discussion of the reason for the decision. With this system, it is relatively easy to find out why a particular position or elevation was or was not used, and if better information becomes available it is easy to change the first decision and re-do the merge. The software also helps resolve position discrepancies by producing KML files that allow display in Google Earth of clusters of discrepant station position estimates.

Handling of alternate station codes is still somewhat rudimentary in this system. The NEIC station book lists over 500 such alternate codes. These are stored in a table and our NEIC and ISC parsers do not create entries in the multi-author SITE table for these codes. However, many of the contributing SITE tables have many alternate codes that are not specifically called out as such. Currently, our system identifies candidate alternate codes by doing a pair-wise comparison of positions for stations in the multi-author SITE table. Candidates not already in the alternate code table are placed in a candidate table where they can be inspected manually. Alternate codes are not included in the final merged SITE table. The current system does not yet handle temporary alternate codes.

We have used our SITE merging system to combine SITE information from the most recent NEIC and ISC station books, the current NNSA GNEMRE SITE tables, and the Incorporated Research Institutions for Seismology (IRIS) SITE table (derived from data-less SEED volumes minus temporary deployments and California stations). There are nearly 36,000 entries in the multi-author site table which produce nearly 14,000 merged SITE entries. There are 166 preferred positions, 41 preferred elevations, 55 rejected positions, and 514 rejected elevations. The position overrides were determined mostly through a combination of inspection in Google Earth and residual analysis using GT events. Most of the elevation overrides were arrived at by comparison of reported elevations with elevations computed using the gtopo30 elevation model.

The GT Merge Effort

It has become necessary to merge the GT25 and better datasets between contributing laboratories for use in both a tomographic inversion for *Pn* velocity of Eurasia and for computing first-*P* correction surfaces using the KBCIT software. The merge is intended both to resolve GT common between labs (choosing the better GT estimate when possible) and to perform an extensive set of quality control steps to the origin and phase data. We have developed a software system implemented in Java, Oracle relational database, and PL/SQL to perform this merge process (Flanagan et al., 2007).

The software brings together into a GTMERGE schema the GT data from both labs along with all supporting ORIGIN, ORIGERR, ASSOC and ARRIVAL data. All data are given new IDs unique within the GTMERGE schema, and events in common between labs are identified by spatial-temporal correlation. The Bondar-Myers-Engdahl-Bergman (BMEB) Epicenter accuracy criteria are used (Bondar et al., 2004). For those events in common, a set of ranking rules is applied to select the best non-BMEB GT. A small subset of the input GT that cannot be ranked is placed in a manual resolution table.

The quality check (QC) steps performed by the software include the following:
- Enforcing common phase naming conventions
- Removing arrivals that are too early or too late to be of interest
- Removing phases not of interest
- Identifying and removing arrivals that are too discrepant to be useful
- Enforcing distance-dependent phase name conventions
- Choosing a “best” arrival for each EVID-STA-PHASE table

After QC is complete, the system evaluates all the BMEB GT for strict adherence to their criteria. All events that fail this check have a new origin solution computed using phase gathers appropriate to the GT level. If the new solution meets its criterion, then it is included in the final merge results. Otherwise, the event is rejected. When all GT have been re-evaluated, a new set of constrained origin solutions is computed using teleseismic P-arrivals. These

"baselined" origins are the final product of the merge effort. The data set produced by the merge effort includes about 97,000 distinct events with nearly 20,000,000 arrivals.

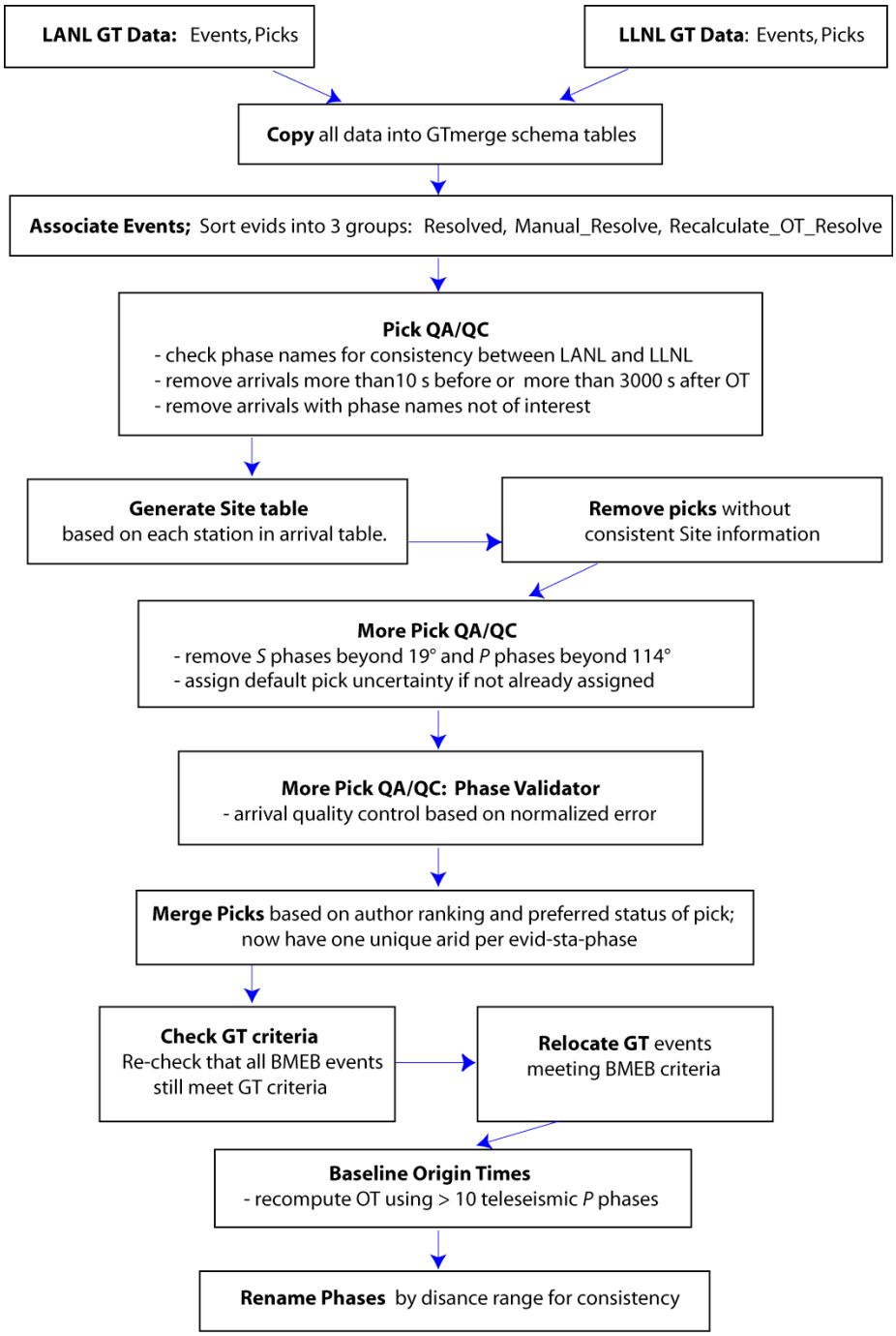

Figure 3. Overview of the GT merge process.

CONCLUSIONS AND RECOMMENDATIONS

We present an overview of our software automation efforts and framework to address the problematic issues of consistent handling of the increasing volume of data, collaborative research efforts and researcher efficiency, and overall reduction of potential errors in the research process. By combining research driven interfaces and workflows

with graphics technologies and a database centric information management system coupled with scalable and extensible cluster based computing, we have begun to leverage a high performance computational framework to provide increased calibration capability. These new software and scientific automation initiatives will directly support our current mission including rapid collection of raw and contextual seismic data used in research, provide efficient interfaces for researchers to measure and analyze data, and provide a framework for research dataset integration. The initiatives will improve time-critical data assimilation and coupled modeling and simulation capabilities necessary to efficiently complete seismic calibration tasks. This GNEMRE Program's scientific automation, engineering and research, will provide the robust hardware, software, and data infrastructure foundation for synergistic calibration efforts.

ACKNOWLEDGEMENTS

We acknowledge the assistance of the LLNL computer support unit in implementing and managing our computational infrastructure. We thank Jennifer Aquilino, Laura Long, and John Hernandez for their assistance in configuration and installation of our Linux cluster and workstations.

REFERENCES

Bondar, I., S. Myers, E. Engdahl, and E. Bergman (2004). Epicenter accuracy based on seismic network criteria, *Geophys. J. Int.* 156: 483–496.

Elliott. A. , D. Dodge, M. Ganzberger, T. Hauk, E. Matzel, and S. Ruppert (2006). Enhancing seismic calibration research through software automation and scientific information management, in *Proceedings of the 28th Seismic Research Review: Ground-Based Nuclear Explosion Monitoring Technologies*, LA-UR-06-5471, Vol. 2, pp. 967–975.

Flanagan, M, D. Dodge, and S. Myers (2007), GT Merge Processing Documentation—Beta Version, in University of California, Radiation Laboratory, 1–15 (in review).

Flanagan, M., E. Matzel, S. van der Lee, H. Bedle, M. Pasyanos, F. Marone, B. Romanowicz, C. Schmid, and A. Rodgers (2006), Joint inversion for three-dimensional velocity structure of the broader Africa-Eurasia collision region, in *Proceedings of the 28th Seismic Research Review: Ground-Based Nuclear Explosion Monitoring Technologies*, LA-UR-06-5471, Vol. 1, pp. 397–406.

Mayeda, K., A. Hofstetter, J. L. O'Boyle, and W. Walter (2003). Stable and transportable regional magnitudes based on coda-derived moment-rate spectra, *BSSA* 93: 224–239.

Ruppert, S., T. Hauk, J. O'Boyle, D. Dodge, and M. Moore (1999). Lawrence Livermore National Laboratory's Middle East and North Africa Research Database, in *Proceedings of the 21st Seismic Research Symposium: Technologies for Monitoring the Comprehensive Nuclear-Test-Ban Treaty*, Vol. 1, pp. 243.