ENHANCING SEISMIC CALIBRATION RESEARCH THROUGH SOFTWARE AUTOMATION AND SCIENTIFIC INFORMATION MANAGEMENT

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

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 (GNEM R&E) Program has automated significant portions of the processes of both seismic data collection and processing, and of determining seismic calibrations and performing scientific data integration by developing state-of the-art tools. We present an overview of our software automation and scientific data management efforts and discuss frameworks to address the problematic issues of very large datasets and varied formats utilized during seismic calibration research. The software and scientific automation initiatives directly support the rapid collection of raw and contextual seismic data used in research, provide efficient graphics-intensive and user-friendly research tools to measure and analyze data, and provide a framework for research dataset integration. The automation also improves the researcher's ability to assemble quality-controlled research products for delivery into the NNSA Knowledge Base (KB). The software and scientific automation tasks provide the robust foundation upon which the synergistic and efficient development of GNEM R&E Program seismic calibration research may be built.

The task of constructing many seismic calibration products is labor intensive and complex, hence expensive. However, certain aspects of calibration product construction are susceptible to automation and future economies. We are applying software and scientific automation to problems within two distinct phases or "tiers" of the seismic calibration process. The first tier involves initial collection of waveform and parameter (bulletin) data that comprise the "raw materials" from which signal travel-time and amplitude correction surfaces are derived, and is highly suited for software automation. The second tier in seismic research content development activities, which we focus on in this paper, includes development of correction surfaces and other calibrations. This second tier is less susceptible to complete automation, more complex and in need of sophisticated interfaces, as these activities require the judgment of scientists skilled in the interpretation of often highly unpredictable event observations. Even partial automation of this second tier, through development of tools to extract observations and make many thousands of scientific measurements, has significantly increased the efficiency of the scientists who construct and validate integrated calibration surfaces. This achieved gain in efficiency and quality control is likely to continue and even accelerate through continued application of information science and scientific automation.

Data volume and calibration research requirements have increased by several orders of magnitude over the past decade. Whereas it was possible for individual researchers to download individual waveforms and make time-consuming measurements event by event in the past, with the terabytes of data available today, a software automation framework must exist to efficiently populate and deliver quality data to the researcher. This framework 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 that is necessary for validation and error analysis. Lack of information management robustness or loss of metadata can lead to incorrect calibration results in addition to increasing the data management burden. To address these issues we have succeeded in automating several aspects of collection, parsing, reconciliation and extraction tasks, individually. We present several software automation tools that have resulted in demonstrated gains in efficiency of producing scientific data products.

OBJECTIVES

The NNSA GNEM R&E 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 GNEM R&E Program calibration efforts. The current task of constructing many seismic calibration products is labor intensive and complex, expensive and error prone. The volume of data and calibration research requirements have 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 are to facilitate development of information products for the GNEM R&E 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 National Data Center run by the Air Force (Figure 1). The calibration processes themselves appear linear beginning with data acquisition (Figure 1) extending through reconciliation, integration, measurement and simulation through to the construction of calibration and run-time parameter products. However, efficient production of calibration products requires extensive synergy and synthesis not only between large datasets and a vast array of data types (Figure 1), but also between measurements and results derived from the different calibration technologies (e.g., location, identification, and detection) (Figure 1). This synergy and synthesis between complex tools and very large datasets is critically dependent on having a scalable and extensible unifying framework. These 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 (Figure 3), using proven research driven workflows and excellent graphics technologies as an 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 (Figure 1). 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 corrections 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. 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 proprietary Java code library, achieving 45% code reuse (in lines of code) throughout several thousand Java classes.

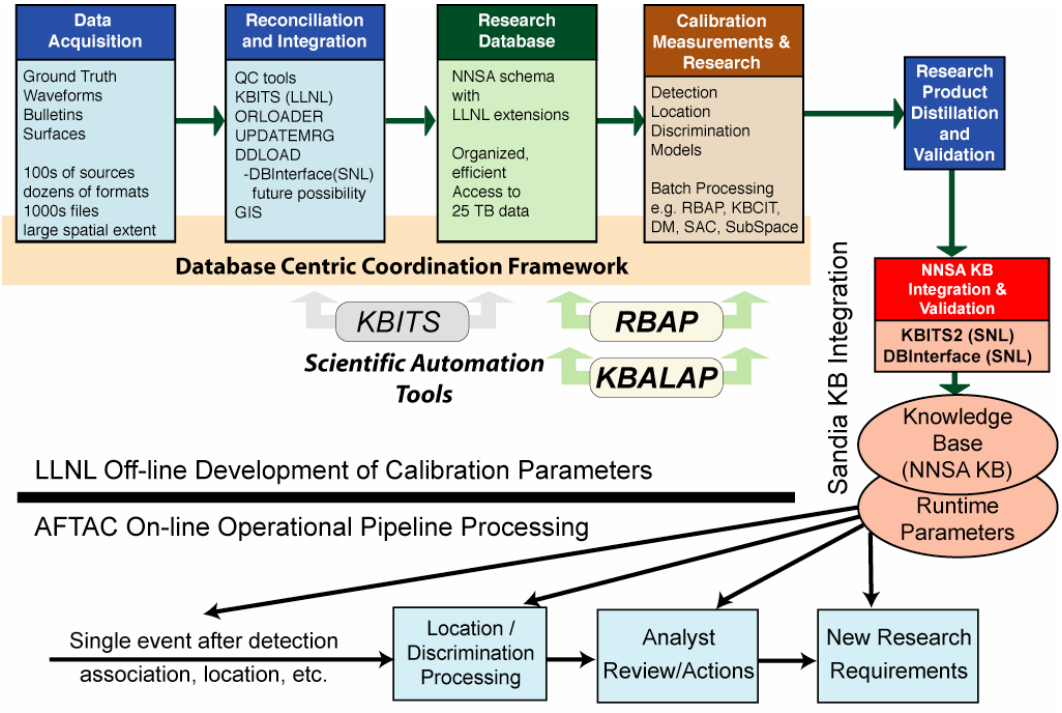

Figure 1: Summary of the processes of data collection, research and integration within the LLNL calibration process that result in contributions to the NNSA KB. The relationships of the current LLNL calibration tools, scientific automation tools, and database coordination framework to those involved in the assembly of the NNSA KB or within the Air Force Technical Applications Center (AFTAC) operational pipeline are delineated.

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 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 (derived measurements such as ground-truth, amplitude measurements, calibration and uncertainty surfaces etc.). Efforts are underway to augment the current relational database structure with structured queries based on semantic graph theory for handling complex queries. 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, PLSQL (extension to Oracle), real-time transactional database triggers, 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.

Sharing of Derived Event Parameters

We have long recognized the inadequacies of the CSS3.0 origin table to serve as a source of information about the "best" parameters for an event. One origin solution may have the best epicenter but poor information on other parameters.

Another may have the correct event type, but be poor in other respects, and so on. We have discussed producing origin table entries with our organization as the author, but that approach has difficulties. Different groups would have responsibility for different fields in the origin. Because their information would not be produced in synchronization, we would either have to always be updating the preferred origin or else producing new preferred origins. Also, there would be difficulties in tracking the metadata associated with each field of the preferred origin. Our solution was to create a set of new tables and associated stored procedures and triggers that collectively maintain the "best" information about events.

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

Automating Tier 1

Corrections and parameters distilled from the calibration database provide needed contributions to the NNSA KB for the Middle East, North Africa and Western Eurasia 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 datasources and external collaborations. Data volumes grew from the 11,400 events with 1 million waveforms in 1998 to the 6 million events with 70 million segmented waveforms and terabytes of continuous data today (e.g. Ruppert et al.; 1999, Ruppert et al. 2005). This rapid increase in stored parameters soon led to two new bottlenecks hindering rapid development and delivery of calibration research.

Automating Tier 2

As the number of 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 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, our development efforts are focused on the development of two scientific automation tools, RBAP and KBALAP, for seismic location and seismic identification calibration tasks, respectively.

Both of these tools include methods and aids for efficiently extracting groups of events and waveforms from the millions contained in the SRDB, and for making large numbers of measurements with metadata in a batch mode. The concept of event sets (groups of related seismic events or parameters that can be processed together, e.g. either station-centric or event-centric) was introduced, as previous seismic analysis code (SAC) scripts and macros could not scale to the task.

All analysis results go directly into the LLNL production schema where they become available for other users. Because users of KBALAP and RBAP may be able to write to our core tables, these tools implement a rule system that uses database roles to control which users can modify data, and which users can modify other users data or modify bulletin data. For example, some users may be able to rank picks, but not save new origin solutions. Some users may be able to perform an array analysis, but not rank picks, etc. By this means we are able to support use of RBAP and KBALAP by analysts with different skill sets and different research priorities. Users get the convenience of being able to produce the results they want and have them immediately available in the production schema without worrying about the impact of their work on a different research group.

The RBAP Program

The Regional Body-wave Amplitude Processor (RBAP) is a station-centric Tier 2 automation tool; it is an interactive, graphical (Figure 2) and highly specifiable software program that acts as a picker and a magnitude and distance amplitude corrections (MDAC) calculator. 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 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 standardized MDAC processing, and it replaces the original collection of LLNL scripts described by Rodgers (2003). 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 integrates the functions of the modules in the previous LLNL scripts into a single, unifying program that is designed to both perform the amplitude measurement task efficiently and to require a minimum effort from the users for managing their data and measurements. For well-located events with pre-existing analyst phase picks, the user reviews for quality control and then generates all the amplitudes with just a few mouse clicks. For events needing more attention, the user has complete control over the process (e.g. window control, ability to mark bad data, define regions, define MDAC parameters and define the events to be used in the overall calibration process). RBAP shortens the time required for the researcher to calibrate each station while simultaneously allowing an increase the number of events that can be efficiently included. RBAP is fully integrated with the LLNL research database. Data is always read directly from the appropriate tables in the research database rather than from a snapshot as was done in the previous system. All RBAP result tables have integrity constraints on the columns with dependencies on data in the LLNL research database. This design makes it very difficult for results produced by RBAP to be stale and also ensures that as the research database expands, RBAP automatically becomes aware of new data that should be processed, as well as data marked as unusable by other applications (such as KBALAP), which should no longer be used for processing. In a like manner, when a segment is marked as bad in RBAP, it is excluded from further processing in KBALAP.

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 other researchers and other projects; interfaces include extensive use of modern mapping technologies and data tables the design of which 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.

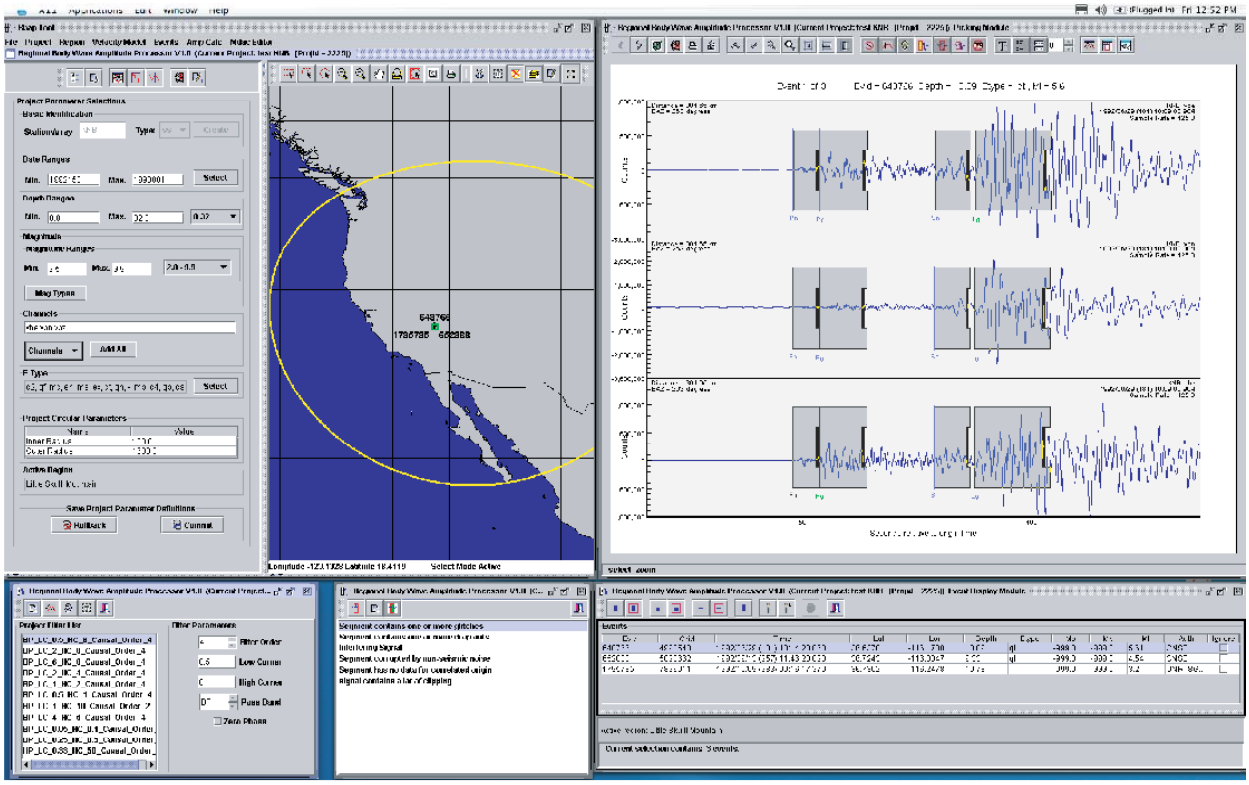

Figure 2: A sample of some of the RBAP graphical and data manager windows.

Some key features of RBAP are listed below:

• *Based on WG 2 Standardized Algorithm*
-RBAP is built on the standardized MDAC body-wave amplitude measurement algorithms. Its results are completely consistent with the last version of the LLNL scripts (Rodgers, 2003) that were vetted in the February 2003 exercise between LLNL, LANL, and AFTAC.

• *Fast and Efficient Calibration*
-RBAP is self-contained and optimized for station-centric body-wave processing. “Good” events can be handled with just few mouse clicks. The researcher has direct control over key calibration parameters within the tool such as phase amplitude windows and migration, marking bad segments, defining distinct geophysical regions, event types to process, etc. We expect RBAP, when fully developed, to provide roughly a factor of 5 increase in calibration speed compared with the original scripts, enabling us to calibrate more stations, with more events per station.

• *Project Management*
- RBAP is designed so that a calibration project can be put down for a day, month or a year, and easily picked up, by the same researcher or a new one. All processing metadata is saved and events are easily tracked as processed, unprocessed or outside the current project definitions. This allows a researcher to efficiently work through a huge data list without repetition and to easily identify and incorporate new events as they become available in the database.

• *Utilizes Database for Up-to-Date Results*
- RBAP can draw on the latest calibration parameters being generated by other working groups, such as the most recent phase picks, relocations, magnitudes, instrument response information, or event type ground truth.

• *Batch Processing*
- RBAP is designed to allow simple batch updating of the amplitude results, whether the change is small (e.g. one-event is relocated) or large (instrument response is changed affecting all events).

• *Engenders Collaboration, Consistency and Efficiency*
- RBAP's complete database integration allows multiple researchers to access the finest-grained tuning parameters for all projects; no data is lost in collaboration, and parameters may be reused.

The KBALAP Program

The KBALAP program is another Tier 2, event-centric automation effort in the GNEM program. It is a highly interactive, graphical tool (Figure 3) 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 ground-truth criteria with no further processing, and for identifying events that would likely meet a predefined ground truth level if a new origin solution was 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, event type,
- batch-mode location of externally-produced GT information,
- production of array azimuth-slowness calibration data, and
- easy review and modification of event parameters used by all GNEM researchers.

Users of KBALAP are able to easily search for data relevant to the production of GT and filter the results by processing status, GT level or potential GT level. The user can select any GT or potential GT event and observe the distribution of stations with picks and stations which have available waveforms. The tool can indicate whether a selected event has the potential to become a GT event if appropriate picks were made on available waveforms that currently have no suitable picks. The user can also select any station with available waveforms and open a picker with any current picks displayed, and adjust existing picks, add new picks, mark bulletin picks as unusable, and relocate the event. When a new GT level is calculated, the user can choose to accept that origin solution and GT level, or continue working with other stations. Traces marked as unusable in KBALAP are automatically viewed by RBAP as bad, and thus not used in processing in that program as well.

With a single mouse click, the user can open a selected event for review and further analysis. In this review mode the user can review and rank existing picks, calculate new origin solutions, and, if appropriate, produce calibration origins. At any point in this process, the user can see the current spatial distribution of arrivals and stations with

waveforms. The tool can also guide the user toward analyzing stations that are important to achieving an origin solution with the best possible GT level.

The interactive GT entry mode of KBALAP allows the user to retrieve information about a specific event and add or update that event's GT parameters. The program can also create a new event with a GT level for cases where epicenter, time, depth and magnitude GT data are available. Similarly, KBALP's batch mode allows for the specification of flat files containing GT data for events already in the database. KBALAP's research driven interface design includes dedicated graphical user interfaces (GUIs) for station filtering, event selection and single and multi-station phase-onset pick windows. Further, there are GUIs that allow users to specify, store and apply different types of bandpass filters. Finally, there is a waveform pick editor window, as well as multiple pick "views," including band filtered views for low signal-to-noise problems and filtering by attributes such as analyst or load date.

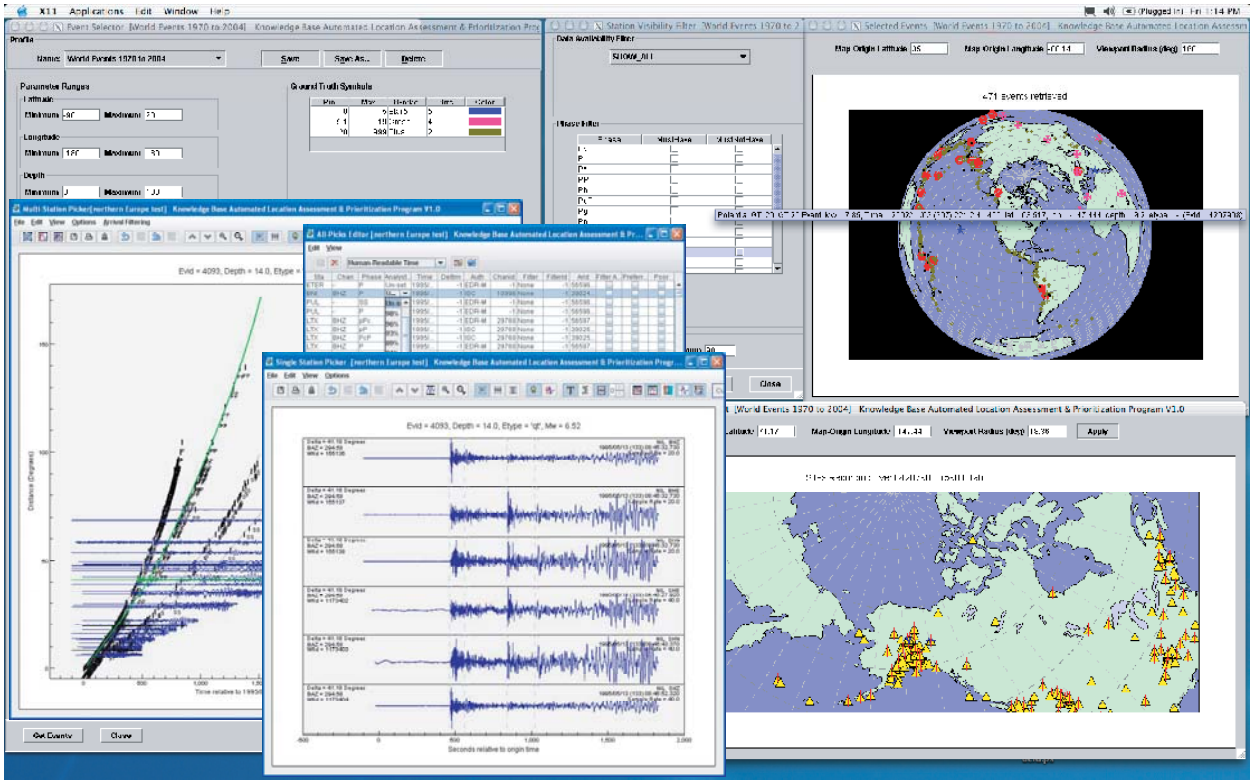

Figure 3: A sample of some of the KBALAP graphical and data manager windows.

Some key KBALAP features are listed below:

• *Fast and Efficient Location*
-KBALAP data selection profiles are self-contained and optimized for the event-centric task of location. KBALAP displays all available picks with available waveforms, and allows picks with waveforms to be modified.

• *Project Management and Collaboration*
- KBALAP is designed so that a profile can be put down for a day, month or a year, and easily picked up by the same researcher or a new one. All processing metadata is saved, and events are easily tracked as processed, unprocessed or outside the current project definitions. This allows researchers or research teams to efficiently work through a huge data list without repetition and to easily identify and locate new events as they become available in the database. Once a senior researcher has reviewed a profile's picks, these picks are finalized and the arrivals and associated metadata made available to other researchers and tools.

• *Batch Processing*
- KBALAP is designed to allow simple batch loading of externally produced GT information.

Further Enhancements to Efficiency Through Cluster-Based Computing

We have begun to leverage scalable and reconfigurable cluster computing resources to improve the efficiency of our computational infrastructure. Just as the database-centric approach to information management provided important gains in efficiency, we have realized the need to move to a different computational paradigm to provide the computational power necessary during calibration production and research. We have begun developing a set of flexible and extensible tools with platform independence that are parallelizable. These research tools will provide an efficient data processing environment for all stages of the calibration workflow, from data acquisition through making measurements to calibration surface preparation. We are also scheduled to implement Oracle 10g's clustering capabilities to further push the performance envelope for our production database. This scalable and extensible approach will result in more coupled and dynamic work flow in contrast to the linear work flow of the past, and allow more interaction between data, model creation, and validation processes.

Initial development and modification of existing codes and algorithms of the cluster based computing environment has yielded significant efficiency improvements in RBAP and other measurement tools. Modification of RBAP to incorporate threads to isolate computationally intensive operations has provided a more interactive and responsive environment for the researcher, as well as laid the ground work for moving the threads to cluster-based computing resources. Other areas under investigation that leverage cluster resources are waveform correlation and subspace detector operations, as well as large-scale event relocations to support the evaluation of ground truth and model calibrations.

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 scientific automation engineering and research will provide the robust hardware, software, and data infrastructure foundation for synergistic GNEM R&E Program calibration efforts.

ACKNOWLEDGEMENTS

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

REFERENCES

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. 234–242.

Ruppert, D. Dodge, A. Elliott, M. Ganzberger, T. Hauk, E. Matzel (2005). Enhancing Seismic Calibration Research Through Software Automation, in *Proceedings of the 27th Seismic Research Review: Ground-Based Nuclear Explosion Monitoring Technologies*, LA-UR-05-6407, Vol. 2, pp. 937-945.