ELECTRONIC SEISMOLOGIST

EMERALD: A Web Application for Seismic Event Data Processing

by John D. West and Matthew J. Fouch

INTRODUCTION

Seismologists studying earthquake sources and the structure of the Earth's interior commonly use event-based seismic data processing, with typical workflow including the acquisition, preprocessing, and analysis of data associated with one or more discrete seismic events. Until recently, most seismic event datasets were relatively small, including at most a few thousands or tens of thousands of seismic waveforms used in a given study (e.g., Dziewonski and Woodhouse, 1987; Grand, 1987, 1994; Woodward and Masters, 1991; Engdahl *et al.*, 1998). More recently, however, the size of datasets has increased significantly, not only from larger-scale investigator-driven temporary field experiments such as the Kaapvaal Project (Carlson *et al.*, 1996), La RISTRA (Wilson *et al.*, 2005), the High Lava Plains Project (Carlson *et al.*, 2005), NorthEast China Extended SeiSmic Array (NECESSArray; Grand *et al.*, 2006), the Carpathians Basin Project (Hetényi *et al.*, 2009), and Project INDEPTH (e.g., Langin *et al.*, 2003; Karplus *et al.*, 2011) but also from large regional networks (i.e., from several dozens to 100+ broadband stations) such as the Advanced National Seismic System (earthquake.usgs.gov/monitoring/anss), Pacific Northwest Seismic Network (www.pnsn.org), Hi-net (http://www.hinet.bosai.go.jp), ORFEUS (http://www.orfeus-eu.org), and AfricaArray (www.africaarray.psu.edu). Furthermore, new projects such as the EarthScope's USArray (www.usarray.org) and SinoProbe (www.sinoprobe.org) are generating unprecedented volumes of data from hundreds to thousands of broadband stations.

The advent of datasets potentially containing millions of seismic waveforms has exposed limitations of traditional seismic processing methods. Preprocessing and processing methods for seismic data vary widely by investigator, but they have some common attributes. They frequently consist of some combinations of standardized command-line programs such as Seismic Analysis Code (SAC; Goldstein *et al.*, 2003) and Generic Mapping Tools (GMT; Wessel and Smith, 1991) and custom modules written or adapted by the investigator in C, FORTRAN, or other programming languages. The standardized and custom modules are generally stitched together using a series of shell scripts that transfer data between modules using a series of flat files. This approach is flexible, is powerful, and has worked well for most natural-source seismic data applications over the years. However, for very large datasets, this approach is much less efficient, and in some applications, this does not work. For instance, shell scripts and operating system calls can break down when confronted with very large numbers of files, the methods for each researcher are dependent on individualized file and directory naming conventions and cannot be easily shared, transferring intermediate data between modules using a series of flat files is inefficient, and basic data processing efforts are duplicated between projects and between researchers. Although some of these issues can be overcome with the application of advanced coding and scripting methods, the skills required for these methods can present a steep learning curve for some new graduate students and significantly limit the ability for new seismic data to be used in simple class projects and other more basic educational settings.

Another issue affecting current dataset handling methods is that once a raw dataset has been acquired from a data center, it has not been possible to simply acquire information regarding updates to seismic station metadata. Station metadata contain, among other things, information regarding precise station location, elevation, sensor orientation, sensor type, and sensor instrument response and are typically acquired along with the initial data download.

To address these limitations of current methodologies, we have developed EMERALD (Explore, Manage, Edit, Reduce, and Analyze Large Datasets), an open-source, easily extensible framework for seismic-event-based processing and analysis. This paper summarizes the present state of the EMERALD system, which is currently in beta testing but is approaching its first formal release. Our beta testers are using the hosted beta version of EMERALD to perform new seismological research and will transition to locally installed copies of the system after formal release. The primary components of EMERALD are outlined here, and we also refer the reader to the online information available on EMERALD's web site at emerald.dtm.ciw.edu.

OVERVIEW OF EMERALD

EMERALD is an open-source web application for downloading, preprocessing, and managing large volumes of seismic event data. A web application is defined as software for which the primary user interface is a web browser. The system need not be exposed to the Internet at large to be considered a web application. Users log-on to an EMERALD server via any

doi: 10.1785/0220110138

modern web browser to request and download or import, preprocess, review, and export seismic data. The web browser interface makes EMERALD independent of the user's computing platform and operating system; as a result, EMERALD can be operated from desktop or laptop computers, tablets, and smart phones. EMERALD always stores the address of the most recently viewed page and defaults to that screen on log-in, enabling users to move seamlessly between computers and/or devices. Nearly all potential users are familiar with the use of a web browser, which reduces the learning curve required and makes new users quickly productive. This feature is especially important for students new to seismic data processing because many may not yet be familiar with command-line-driven data analysis methods.

EMERALD includes an integrated data-request module through which seismic event data can be requested from the Incorporated Research Institutions for Seismology Data Management Center (IRIS DMC; www.iris.edu/dms/dmc) and imported directly into the EMERALD database. In addition, seismic data retrieved via other methods including SOD (Owens *et al.*, 2004), JWEED (www.iris.edu/manuals/jweed), and BREQ_FAST (www.iris.edu/manuals/breq_fast.htm) can be imported into EMERALD from a collection of SAC files. Importing routines for other file formats (e.g., SEED, MiniSEED, SEISAN, GSE, SEGY, etc.) will be included in future releases depending on community demand and implementer availability. EMERALD includes many tools for basic data preprocessing such as filtering, trimming, demeaning, and rotating seismograms; calculating estimated phase arrival times; eliminating duplicate, incomplete, or noisy traces; and rapidly viewing and accepting or rejecting large numbers of traces. Users can move easily between event-centric and station-centric views of their data. For example, a user trace editing in event view can switch to a station view for a particularly noisy station to determine whether the entire output of that station should be eliminated then return back to event view to continue trace editing by event. At the completion of preprocessing, the resulting edited dataset can be exported from EMERALD for further analysis.

THE DATASET

A central concept in EMERALD is that of the dataset. In EMERALD parlance, the dataset is the complete set of seismic event time series, event and station metadata, time-series metadata (sampling rate, start and end times, etc.), and calculated or parametric data required for a given seismic project or investigation. The home page for an EMERALD dataset provides an overview snapshot of dataset parameters, including a map of sources and receivers, counts of waveforms, and histograms of data by year, magnitude, and back azimuth (Fig. 1). EMERALD calculates summary information by dataset, and calculations and processes can easily be applied to all traces in a dataset. A dataset can consist of one or more subsets, each of which is derived from some portion of the original data by some combination of processes such as filtering, trimming, rotating, etc. The user controls the number of processes applied to each subset, striking a balance between the flexibility of maintaining a full set of waveforms at each step with the disk space needed to do so. Thus, the data in a sequence of subsets represent various checkpoints in the processing of the dataset and can be returned to by the user for subsequent reprocessing without having to restart a workflow from raw data. Each processing step is also logged, so processing history and process parameters can be viewed later.

DATA-REQUEST MODULE

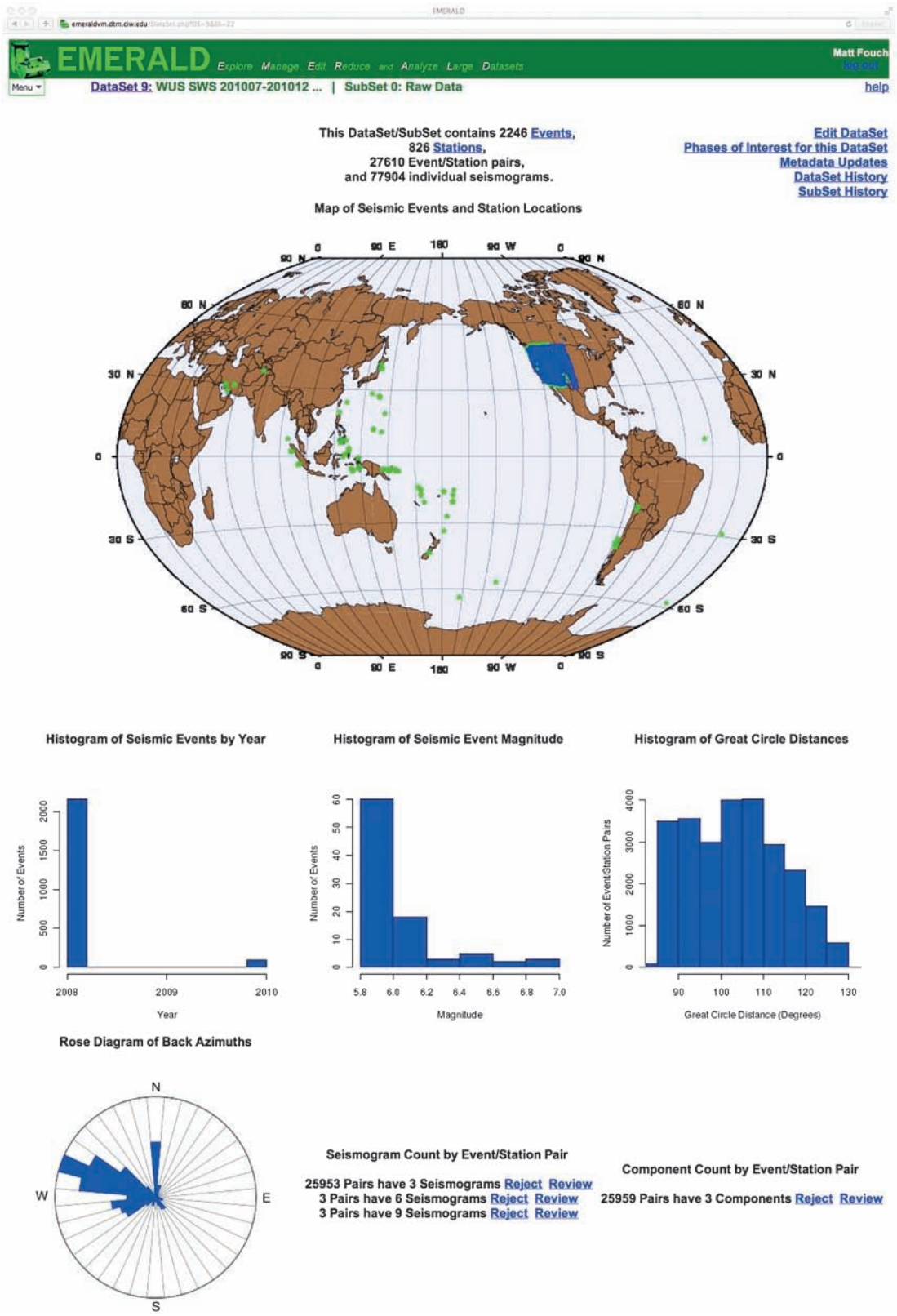
Users can directly request data from the IRIS DMC using the EMERALD request module. Within the data-request module, the user defines a range of dates, a time window (usually some number of seconds before and after a particular seismic phase), and a group of channels of interest. Stations can be selected from a comprehensive list, by network, or within a rectangular, circular, or annular geographic area. Stations can also be specified to reside within a given angular distance range from each event. Similarly, events are specified to have a given magnitude and depth range and can be specified to be within a rectangular, circular, or annular geographic area or to be within a given angular distance from each station.

Station and event lists are populated in the database from the IRIS DMC ws-station and ws-event web services (www.iris.edu/ws) and are kept up-to-date by background processes. The event catalogs are derived from the ANF, GCMT, ISC, and NEIC PDE catalogs and can be selected by individual catalog or from a combination of catalogs. There is currently no facility for importing user-defined catalogs, but this is a feature that could be added in a future release.

The requesting module preprocesses the request by matching events to stations based on date, time, and phase or phases requested and calculating a time window based on the event time and the expected phase arrival time. Preprocessed requests are bundled by event and requested from the IRIS DMC using the ws-bulkdataselect web service. The returned data are read directly into the EMERALD database without user intervention. Although waveform (time-series) data and time-series metadata can only be requested via the IRIS DMC in the current version, EMERALD includes the programming interfaces to request data and metadata from other sources, and therefore, connections to other data centers can be added as indicated by the user community. A first likely candidate for additional data center requesting capability is the European Earthquake Data Portal (www.seismicportal.eu), which has implemented web services for data and metadata compatible with EMERALD.

TRACE EDITING MODULE

A core feature of EMERALD is the capability for quickly reviewing large numbers of seismic traces. Users can scroll through a web page displaying all traces for a given event–channel combination (Fig. 2) or a similar page displaying all traces for a given station–channel combination. A related page



▲ **Figure 1.** Home page for a dataset. This gives the user a general overview of the data by tabulating the number of events, stations, and individual seismogram. A world map shows the distribution of events and stations, a set of histograms give the distribution of events by year, magnitude, and great circle distances, and a rose diagram shows the back-azimuth distribution. Event–station pairs with more or fewer seismograms and/or components than expected can be easily reviewed or batch rejected from the links on this page. The home page also has handy links for editing the dataset properties, setting up metadata updating, and reviewing the history of the current dataset and subset.

▲ **Figure 2.** Reviewing traces by event. For any chosen event, all traces for a selected seismic component (channel or orthogonal axis) can be viewed and accepted or rejected. Here, all radial-component traces are shown for event 2010.355.17.19.40. The ability to quickly scroll through the traces and accept or reject individual seismograms or all traces for a specific event or station is one of the key methods provided by EMERALD for preprocessing large volumes of seismic data.

displays all traces for a given event–station combination. Checkboxes provide the means for accepting or rejecting individual traces, all traces for an event, or all traces for a station. In addition, in the event–station view, users can grade the acceptability of the traces for that event–station combination on a 0–5 scale. The ability to view and reject all data from a malfunctioning station or noisy event with one click greatly contributes to the ease of generating a more useful seismic dataset for subsequent analysis.

METADATA UPDATER

As mentioned earlier, a persistent challenge faced by seismic researchers has been the issue of seismological station metadata that change following data download for analysis, or coupling of new data with obsolete metadata acquired at an earlier time. Station metadata describe characteristics of the station and/or instrumentation installed at that station, such as location, elevation, depth, azimuthal orientation of horizontal components, instrument response, station name, network code, etc. Elements of these metadata may change over time for a number of reasons. A few examples are as follows: instruments may fail and be repaired or replaced (thus changing the instrument type and impulse response information), errors in station location or instrument azimuthal orientation may be discovered and corrected, or the station may change ownership and be renamed to conform to the new operator's network naming convention.

Prior to EMERALD, there was no available method for detecting changes in station metadata and reporting these to the investigator. Researchers typically either acquire metadata at the same time data are acquired from a centralized data management center or rely on already-acquired metadata when downloading new waveforms. Subsequent metadata changes, therefore, are unknown and cannot be addressed by the investigator without developing a special set of scripts or code that track this issue. Occasionally, significant problems with station metadata lead to a query of metadata, but in the event of changes, manual review and modification of files is the most common approach. This manual method of review and modification is time consuming and tedious for even a few stations, and with the advent of datasets from thousands of seismic stations, it becomes time prohibitive and is thus usually neglected, although some of those updates can affect significantly the results of a given investigation.

EMERALD includes capabilities to store station metadata, check for updated metadata as a periodic background process, compare newly acquired metadata with previously stored snapshots, and notify the user of metadata changes. Users can specify, on a project-by-project basis, which classes of metadata are of importance and how they would like changes handled. Metadata changes can be automatically applied to the dataset or held for approval, and users can choose whether to apply metadata updates with or without notification to and approval by the user. When acquiring new seismograms, metadata are matched to the seismograms based on the date and time of the seismic trace, ensuring that the most current metadata are always associated with a seismic trace.

DATA EXPORT MODULE

At the completion of preprocessing, EMERALD provides methods for the user to export the clean dataset as a set of SAC files for further processing. To maximize compatibility with external processing routines, SAC file exports from EMERALD can be organized in directory structures by event or by station, and the user has control of the naming convention used for exported files. Although the export module currently is implemented for SAC files only, exporting of additional file formats is planned for future releases of the system.

WORKFLOW FEATURES

Workflow automation, at its most basic, is the serialization of a number of data processing steps so that they are executed one at a time. More sophisticated workflow schemes add the functionality for branching instruction sets, error detection and/or correction, and the capability to restart the batch from chosen points in the workflow process. Most workflow systems include logging of each step along with any error messages or notifications.

In this initial version of EMERALD, we provide users the capability to assemble multiple processing steps into a workflow, termed an "automation batch." Such workflows are stored and available for repeated use, so multiple datasets can be processed identically. A typical workflow might include such steps as calculating estimated arrival times, trimming traces to a specified window around a phase arrival, removing the mean, bandpass filtering, rotating from ENZ (east/north/vertical) to RTZ (radial/transverse/vertical) coordinate systems, and calculating (using one of several methods) the signal-to-noise ratio for each trace. Large datasets and complex workflows can yield relatively lengthy processes, so EMERALD includes a notification feature that alerts the user via e-mail or text message of the completion of the workflow and optionally at the completion of each step. All processing in EMERALD, including processing within an automation batch, is logged to the EMERALD database for easy review by the user.

INTERNAL DESIGN

EMERALD is a complete virtual appliance composed of an Ubuntu Linux operating system (www.ubuntu.com), Apache web server (httpd.apache.org), PostgreSQL Relational Database Management System (www.postgresql.org), active web pages constructed using the PHP scripting language (www.php.net), and all required drivers and libraries. All included software is free and/or open source, and all code written by EMERALD developers is freeware and thus can be modified in any way by the user. Users modifying a particular feature can lock out that feature so that automatic updates do not overwrite the user's changes. The system supports extensions written in a wide range of programming languages and incorporates many existing seismic data processing tools, including TauP (Crotwell *et al.*, 1999), SeisFile (Owens *et al.*, 2004), GMT (Wessel and Smith, 1991), and ObsPy (Beyreuther *et al.*, 2010). The PostgreSQL database management system was chosen because it is a mature, high-performance system that allows extensions to be written in a wide range of languages and because it incorporates built-in array data types, allowing seismic time series to be stored in the database as arrays of double-precision numbers. Additional information regarding the database schema is beyond the scope of this article but will be described in detail in a future manuscript about EMERALD design, development, and internal structure.

The virtual appliance is supplied to the user as a single-disk image file, which can be easily installed using any of the widely available virtualization platforms (hypervisors), such as VirtualBox (www.virtualbox.org), VMware (www.vmware.com), and Xen (xen.org). Many of these hypervisors are free and/or open source. One feature of hypervisors and virtual appliances is that virtual disks can dynamically expand up to a maximum size. Thus, the EMERALD drive can be defined as 2TB (the current maximum individual disk), but the actual drive space may be much smaller, depending on the size of the project.

EMERALD is designed to be installed on a server or desktop workstation for use by an individual researcher or small workgroup. For performance reasons, large research groups may need to have multiple installations or utilize a larger server. Long-term temporary accounts are available on the EMERALD test server at the Carnegie Institution of Washington to allow users to gain experience with the software before committing to hardware for their own EMERALD system. A user's database on any EMERALD system can be exported to a single file for transfer to any other EMERALD system, so work done on the Carnegie server can be moved easily to the user's personal system once they have adopted EMERALD.

EXTENSIBILITY

EMERALD is easily extensible, and new features can be developed in any of a wide range of programming languages. Languages with native access to the EMERALD database include C, Java, Python, Perl, PHP, R, Ruby, Tcl, and Lua. Other languages are being ported to the PostgreSQL database system and

will be available in the future. The standardized database schema allows code developed on any user's copy of EMERALD to be installed and run on any other user's copy. Application menu items are stored in database tables, making newly installed methods instantly available via the EMERALD menu structure. Existing command-line tools can be supported through the practice of writing the data to be processed from EMERALD to a virtual disk, calling the function via execution of a shell command, and then reading back the results. This is the methodology currently used to integrate GMT (Wessel and Smith, 1991) into EMERALD, although new programming interfaces to GMT in C and Python will shortly be available (Wessel *et al.*, 2011) and will facilitate future implementation of GMT mapping and plotting methods in EMERALD.

Concurrent with the initial release of the software, all EMERALD source code will be published on the new IRIS SeisCode repository (seiscode.iris.washington.edu), and a system for discovery and download of new and updated modules will be maintained on the EMERALD web site. Community members contributing new modules will submit them via the EMERALD web site, where they will be reviewed by volunteers (primarily to prevent inclusion of malicious code) before being made available to users.

Instructions for user development of new add-on modules for EMERALD are beyond the scope of this paper and are subject to change as the system is modified in the ongoing development. As mentioned earlier in the Internal Design section, a future publication will describe EMERALD's internal structure and the methodology for developing add-on modules, and add-on module templates will be available on the EMERALD web site. Those wishing to create add-on modules in the interim period are encouraged to contact the corresponding author directly for further information.

PROJECT STATUS

EMERALD is an ongoing research project, currently being beta tested by a group of ~25 researchers, including students, postdoctoral researchers, and faculty from a range of scientific and educational institutions. Planned improvements to the system are being driven primarily by feature requests from the beta users. Upcoming new features include the following:

- A centralized site for add-on modules and updates, with automated methods so that individual users can detect, download, and install updates, database modifications, and new methods from within EMERALD.
- Methods for handling synthetic seismograms and for associating synthetics with real seismic data traces.
- Methods to acquire time-series data and metadata from data centers in addition to the IRIS DMC.
- Methods to easily flip (multiply amplitudes by −1.0) individual traces or all traces for a particular station and component combination.
- Methods to manually pick arrivals and/or processing windows.
- Improved scientific workflow features, allowing users to create branching workflows and to easily save, export, share, and import workflow plans.
- A queuing system to provide increased efficiency by spreading data processes across all available processors and cores on the user's server.

During the current beta-testing period, EMERALD is hosted on a centralized server at the Carnegie Institution of Washington's Department of Terrestrial Magnetism to facilitate frequent updates to the system. Those wishing to participate in the beta-testing program are encouraged to contact the corresponding author.

Additional up-to-date details on the status of EMERALD along with screen shots and instructions for use can be found on the EMERALD web site. Feedback, bug reports, and requests for new features can be found on the EMERALD online discussion board at emerald.dtm.ciw.edu/board. This is a closed-access discussion board; contact the corresponding author for access.

CONCLUSIONS

EMERALD provides significant advantages relative to most existing methodologies for seismic data management and processing. The PostgreSQL database engine allows EMERALD to easily handle datasets larger than 1 million records; the data-request module enables efficient requesting and download of waveform data; the simple, intuitive, graphical user interface speeds data review and accelerates the learning curve for new users; and the system checks for updated station metadata and alerts the user to changes. The standardized database format and plug-in architecture provides a mechanism for easy exchange of processing methods between researchers. We hope for broad community adoption of the EMERALD framework and encourage comments and suggestions for its improvement.

ACKNOWLEDGMENTS

No project of this scope and size can be possibly realized without the input and support of a large cast of characters. We gratefully acknowledge the assistance and contributions of the staff at the Incorporated Research Institutions for Seismology Data Management Center (IRIS DMC), including Chad Trabant, Tim Ahern, Bruce Weertman, Alex Hutko, Manoch Bahavar, Rich Karstens, Yazan Suleiman, and Linus Kamb. We appreciate the hardware and systems support from Mark Stevens at the Arizona State University's School of Earth and Space Exploration and Michael Acierno at the Carnegie Institution of Washington's Department of Terrestrial Magnetism (DTM). Partial financial assistance was provided by DTM and IRIS. We particularly acknowledge the ideas, interest, and support of the entire EMERALD users group, especially members Nick Schmerr, Mike Thorne, Chin-Wu Chen, Colton Lynner, Erin Wirth, Joel Derig, Maggie Benoit, Kelsey Brunner, Eli Raymond, Meghan Miller, and

Peiying Lin. We also appreciate the useful advice from Ramón Arrowsmith, Philip Crotwell, and Gary Pavlis.

REFERENCES

Beyreuther, M., R. Barsch, L. Krischer, T. Megies, Y. Behr, and J. Wassermann (2010). ObsPy: A python toolbox for seismology, *Seismol. Res. Lett.* **81,** no. 3, 530–533.

Carlson, R. W., T. L. Grove, M. J. de Wit, and J. J. Gurney (1996). Anatomy of an Archean craton: A program for interdisciplinary studies of the Kaapvaal craton, southern Africa, *Eos Trans. AGU* **77,** 273–277.

Carlson, R. W., D. E. James, M. J. Fouch, T. L. Grove, W. K. Hart, A. L. Grunder, R. A. Duncan, G. R. Keller, S. H. Harder, and C. R. Kincaid (2005). On the cause of voluminous magmatism in the northwestern United States, *Geol. Soc. Am. Abstr. Progr.* **37,** no. 7, 125.

Crotwell, H. P., T. J. Owens, and J. Ritsema (1999). The TauP toolkit: Flexible seismic travel-time and ray-path utilities, *Seismol. Res. Lett.* **70,** 154–160.

Dziewonski, A. M., and J. H. Woodhouse (1987). Global images of the earth's interior, *Science* **236,** 37–48.

Engdahl, E. R., R. van der Hilst, and R. Buland (1998). Global teleseismic earthquake relocation with improved travel times and procedures for depth determination, *Bull. Seismol. Soc. Am.* **88,** no. 3, 722–743.

Goldstein, P., D. Dodge, M. Firpo, and L. Minner (2003). SAC2000: Signal processing and analysis tools for seismologists and engineers, in *The IASPEI International Handbook of Earthquake and Engineering Seismology*, W. H. K. Lee, H. Kanamori, P. C. Jennings, and C. Kisslinger (Editors), Academic Press, London.

Grand, S. P. (1987). Tomographic inversion for shear velocity beneath the North American plate, *J. Geophys. Res.* **92,** 14,065–14,090.

Grand, S. P. (1994). Mantle shear structure beneath the Americas and surrounding oceans, *J. Geophys. Res.* **99,** 11,591–11,621.

Grand, S., Y. Chen, H. Kawakatsu, Q. Chen, J. Ni, F. Niu, M. Obayashi, and S. Tanaka (2006). NorthEast China Extended SeiSmic Array (NECESSArray): Deep subduction, mantle dynamics, and continental evolution beneath northeast China, in *2006 Western Pacific Geophysics Meeting Abstracts*, American Geophysical Union, Washington, D.C.

Hetényi, G., G. W. Stuart, G. A. Houseman, F. Horváth, E. Hegedűs, and E. Brückl (2009). Anomalously deep mantle transition zone below Central Europe: Evidence of lithospheric instability, *Geophys. Res. Lett.* **36,** L21307, doi: 10.1029/2009GL040171.

Karplus, M. S., W. Zhao, S. L. Klemperer, Z. Wu, J. Mechie, D. Shi, L. D. Brown, and C. Chen (2011). Injection of Tibetan crust beneath the south Quaidam basin: Evidence from INDEPTH IV wide-angle seismic data, *J. Geophys. Res.* **116,** B07301, doi: 10.1029/2010JB007911.

Langin, W. R., L. D. Brown, and E. A. Sandvol (2003). Seismicity of Central Tibet from project INDEPTH III seismic recordings, *Bull. Seismol. Soc. Am.* **93,** 2,146–2,159.

Owens, T. J., H. P. Crotwell, C. Groves, and P. Oliver-Paul (2004). SOD: Standing order for data, *Seismol. Res. Lett.* **75,** 515–520.

Wessel, P., and W. H. F. Smith (1991). Free software helps map and display data, *Eos Trans. AGU* **72,** 441.

Wessel, P., W. H. Smith, R. Scharroo, and J. M. Luis (2011). The Generic Mapping Tools (GMT) Version 5, *AGU*, Abstract #IN21D-05 (Fall Meet.), San Francisco, CA.

Wilson, D., R. Aster, M. West, J. Ni, S. Grand, W. Gao, W. Baldridge, and S. Semken (2005). Lithospheric structure of the Rio Grande rift, *Nature* **433,** no. 7,028, 851–855, doi: 10.1038/nature03297.

Woodward, R. L., and G. Masters (1991). Lower-mantle structure from *ScS–S* differential travel times, *Nature* **372,** 231–233.

John D. West
Matthew J. Fouch[1]
School of Earth and Space Exploration
Arizona State University
PO Box 871404
Tempe, Arizona 85287-1404 U.S.A.
john.d.west@asu.edu

[1] Now at Department of Terrestrial Magnetism, Carnegie Institution of Washington, 5241 Broad Branch Road, NW Washington, D.C. 20015-1305 U.S.A.