A Distributed Approach to Computational Earthquake Science: Opportunities, Challenges, and New Approaches

Andrea Donnellan, Jay Parker, Margaret Glasscoe

Jet Propulsion Laboratory, California Institute of Technology

Marlon Pierce and Geoffrey Fox

Indiana University

Dennis McLeod

University of Southern California

Donald Atwood

Alaska Satellite Facility

*Abstract*— Summary, challenges, opportunities, new approaches.

Table of Contents

1. Introduction 1

2. Data Storage and Access 2

3. Computational Infrastructure 4

4. Interfaces to External Applications 5

5. Visualization 5

6. Science Problems as a Driver 2

7. Conclusions 6

Acknowledgements 6

References 6

Biographies Error! Bookmark not defined.

1. Introduction

Earthquake science depends greatly on numerous data types spanning spatial scales from microscopic to global and timescales of fractions of seconds to millions of years making it a rich environment for the application of cyberinfrastructure. As the inadequate preparation and response to recent major earthquakes in Haiti, Chile, and Japan have shown, the field is ripe for transformation: formerly isolated groups must work more effectively with each other. Data providers need to better understand how their data are consumed and fused with other data sources by downstream geophysicists. Geophysicists must understand how to relate their work to emergency planners and responders. Experts focused on the processes of particular areas of the globe must find ways to translate their knowledge to other regions and other research teams. All must be focused on identifying and tackling grand challenges that span areas of expertise. Collaboration alone is not enough: the field needs a common framework designed to foster the desired connections. This is especially imperative as datasets and sources grow and as new spaceborne missions and ground based networks contribute to the field.

Numerous and growing online seismic, geologic, and geodetic data sources from government agencies and other resources around the world provide an exceptional opportunity to integrate varied data sources to improve our understanding of earthquakes. The data support comprehensive efforts in data mining, analysis, simulation, and forecasting. However, the uncoordinated but improving state of current data collections, robustness of data repositories, and the lack of formal modeling tools capable of ingesting multiple data types hamper earthquake research activities. A growing number of research groups and communities are recognizing the need to integrate the heterogeneous data and models. Such groups include, but are not limited, to the Asia Pacific Economic Cooperation (APEC) Cooperation for Earthquake Simulation (ACES), The Alaska Satellite Facility (ASF), Computational Infrastructure for Geodynamics (CIG), EarthScope, the Southern California Earthquake Center, the Scripps Orbit and Permanent Array Center (SOPAC), the University NAVSTAR Consortium (UNAVCO), and E-DECIDER. Interactions between QuakeSim, ACES, and ASF illustrate how improved computational infrastructure facilitates integration of data and models and also highlights new challenges.

NASA’s QuakeSim project focuses on computational infrastructure for modeling and mining remotely sensed and other earthquake related data. QuakeSim development has resulted in a number of successes but has also identified a number of key challenges related to data, computational infrastructure, and infrastructure for modeling, analysis, and visualization. ACES is a coordinated international effort linking complementary nationally based programs, centers, and research teams focused on earthquake researchh. ACES aims to develop realistic supercomputer simulation models for the complete earthquake generation process, thus providing a "virtual laboratory" to probe earthquake behavior. Such capability provides a powerful means to study the earthquake cycle, and hence, offers a new opportunity to gain an understanding of the complete earthquake cycle including earthquake nucleation and precursory phenomena. The complexity of phenomena and range of scales from microscopic to global involved in the earthquake generation process make understanding earthquakes a grand challenge. The Alaska Satellite Facility (ASF) downlinks, archives, and distributes satellite data. Its mission is to promote, facilitate, and participate in the advancement of remote sensing in order to support national and international Earth science research, field operations, and commercial remote-sensing applications that benefit society. ASF recognizes the need to provide high quality data and services in a timely manner.

QuakeSim, ACES, and ASF have identified key challenges as they strive to integrate models, data, and applications for improved science analysis. Such challenges include the need for more open processes, greater integration and accountability of different groups on each other, and managing data and interfaces to overcome bandwidth limitations. The solutions are partially technical and partially sociological. Our focus here is on the interseismic part of the earthquake cycle that is addressed by remotely sensed crustal deformation data, but the challenges and opportunities extend to other data types and parts of the earthquake cycle. An essential driver of these requirements is the need to support both science and disaster response and planning user communities.

Modern technologies such as cloud computing, workflow and (web) services provide an opportunity to address these challenges and require a comprehensive set of activities. These include 1) developing robust bridging services in a service-oriented architecture to integrate data from multiple sources (Figure 1); 2) Developing a fundamental framework for model optimization through the integration of multiple data types; 3) Developing cyberinfrastructure within science gateways to handle the computing requirements of the optimization framework including the need to access large datasets; 4) Ensuring data handling issues of model contribution, provenance, version tracking, commenting, rating, etc. and; 5) Developing capabilities for using output in downstream applications.

2. Science Problems as a Driver

Basic and applied science tasks illustrate diverse needs for coupling tools and data. Carrying out science problems within a computational environment can help to identify issues and is effective for further development of tools and functionality. There must be an adequate number of easily accessible tools available to make this appealing to scientists and end-users such as emergency responders. End-users have lower tolerance for tools that don’t immediately address their needs and may be less inclined to adopt their usage if they don’t work well at the outset. The ideal scenario is to develop the tools with friendly users and then further develop documentation and expand the user base as the tools become more functional.

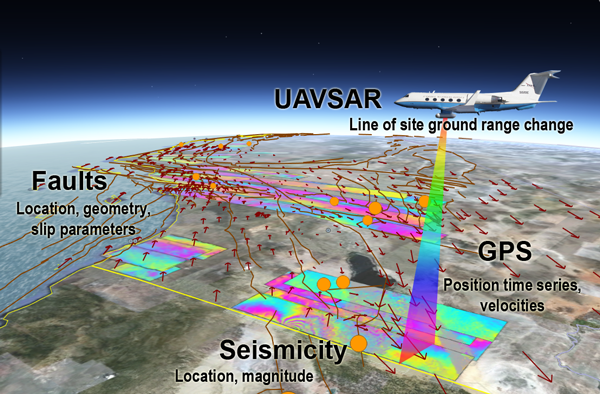


Figure 1. Example data sources

Science use cases come with a different set of challenges compared to end users such as emergency responders who need a well-developed set of tools that can be used routinely. Science studies often deviate from routine tool usage. Part of the scientific process involves exploring data or models in new ways. As a result keeping up with new tools to satisfy ever changing scientific approaches is challenging. Scientists need toolboxes to develop new approaches more than standardized workflow and the infrastructure and personnel needed to develop toolboxes that allow for flexible analysis of the data is quite extensive. It is important to identify routine tasks that scientists typically carry out during data or model exploration and create tools that minimize duplication of effort in order to add efficiency to the scientific process. Use cases grouped under three modes of science understanding, forecasting, and response, illustrate different user needs and the potential interaction with computational infrastructure and tool developers.

*Use case: Science Understanding*: Scientist identifies regions of active crustal deformation from GPS and InSAR/UAVSAR data products. GPS products can be in the form of position time series or station velocities. The scientist scans through the velocity data plotted in vector form on a map in different reference frames to guide thinking as to where active crustal deformation is occurring. The scientist inverts crustal deformation data for fault motions constrained by paleoseismic fault data and then develops simulations based on fault locations and behavior. The scientist may search GPS time series for transient anomalies that indicate previously unknown characteristics of crustal behavior. The possibilities are numerous and the scientist generally wants to explore the data in new ways. Many steps are routine, however, and a scientific user affiliated with computational tool developers can assess which tasks are carried out frequently enough to warrant new tool development.

*Use case: Forecasting*: Scientist identifies active faults from multiple data sources: GPS, UAVSAR, InSAR, paleoseismic fault data, seismicity. This is likely to be an outgrowth of the above scientific understanding and exploration. Once techniques are developed pattern analysis is carried out to search for anomalous features in GPS time series and seismicity data. Interacting faults are simulated and statistical analysis of the interactions is conducted. Earthquake probabilities are evaluated for short to decade time scales. Ultimately these probabilities are integrated into the Working Group on California Earthquake Probabilities (WGCEP) Uniform California Earthquake Rupture Forecast (UCERF). The analysis techniques must be well understood and well defined or standardized in order to incorporate the probabilities into UCERF, which is an official earthquake probability model published by the US Geological Survey [4].

*Use case: Response*: When an event occurs deformation can initially be estimated from models that use available seismic information. Initially that information is location, depth, and magnitude of the event. As a result assumptions must be made about the possible mechanism. Where fault data are available the likely mechanism can be constrained to known faults. In time an earthquake mechanism is produced, which provides two orthogonal geometries of slip. The deformation estimates can be used to estimate the envelope of maximum displacement, and hence most likely region of damage. This envelope can also be used to guide acquisition of UAVSAR and GPS data for emergency and science response. Possible locations of future aftershocks can be assessed, as the fault models are refined. The damage zone from the event can be defined as a polygon and formatted for ingestion into loss estimation tools. Damage and potential aftershock assessments can be refined as new information becomes available. The products can be made available to emergency responders. Such products must be easily accessible and intuitively interpretable by responders. It is highly important that any tools used be mature and support the disaster community as well as sophisticated science users.

3. Data Storage and Access

Integrating and modeling the ever-growing and increasingly multisource geodetic GPS and Interferometric Synthetic Aperture Radar (InSAR) data volumes is necessary to improve crustal deformation models. The models are used for forecasting, simulation, emergency planning and response applications. Remotely sensed data provide estimates of crustal deformation that are key to improving fault models. GPS data provide long-term estimates of crustal deformation of a network of California and global sites. A time series of the daily change in position of these sites provides detailed information about temporal crustal changes. Current InSAR data products provide detailed images and spatial distribution of crustal deformation sparsely sampled in time. A spaceborne InSAR mission dedicated to studying surface deformation would provide routine high-resolution interferograms of ground motions adding a significant increase in the temporal and spatial resolution of InSAR data products. Even so, airborne UAVSAR data products are large and complex enough that maximizing their utility is best done when tools are developed that take into account bandwidth and different user groups.

Data storage, processing, mining, and analysis challenges need to be addressed now to maximize the utility of earthquake related missions and projects [1]. Design, launch, and operation of space and airborne missions represent a significant cost. The benefits of such investments will only be realized if the infrastructure is in place for investigators and users to access and interpret those data.

Efficiently analyzing, integrating, and modeling geodetic and geologic data requires digital storage of the data, including the fault specifications, and automated access to the data through network services. As the data sources, volumes and regions of interest grow it is necessary for applications, not just humans, to access the data for remote automated processing. The data are distributed and under the cognizance of a wide array of agencies and institutions. Developing standards through formal and informal collaborations and partnerships is key to maximizing the use of solid Earth science data. Numerous processes result in deformation of the Earth’s surface. Accessing, mining, and modeling crustal data are key to understanding these processes. The potential applications of Earth surface data, such as simulations, are varied, include a large globally distributed set of users, and create archives many-fold times larger than the centers that process or store the data.



Figure 2. QuakeTables database showing UAVSAR map browser and links to data products.

Models require an increasing number of types of data to guide them. The data are of many different forms and sizes. Fault data, for example, yield information about fault geometry, slip rates, and earthquake recurrence. At the other end of the spectrum interferometric radar data tend to be in large binary image files on the order of 1 GB/image. QuakeSim applications use data products, rather than the raw data. The Alaska Satellite Facility (ASF) Distributed Active Archive Center (DAAC) is NASA’s assigned facility for UAVSAR and InSAR data and as such both ASF and QuakeSim must taking into account the necessary interfaces between the data products and applications. QuakeSim applications use data from many sources including: fault data, GPS time series and velocities, seismicity data, seismic moment tensor information, and interferometric synthetic aperture radar (InSAR) images.

Understanding the origin and processing of the data products is important for assessing their quality for ingestion into models. Data products often change with time as new processing techniques or new interpretations become available. One key challenge is keeping up with and documenting improved data products from newer processing techniques as they become available. Ideally, there is a feedback loop from modelers to data product providers that enables modelers to identify issues with the data and request reprocessing of data. This feedback loop will result in greater utility of the data than training modelers to process complex data as well as model and interpret the results. Both data processing techniques and model development are so complicated that they can take careers to develop and as a result teams of people with varied roles rather than individuals must contribute to the final analysis.

Data products, even for the same data type, are not standardized and are often not adapted for machine interfaces. This requires manual input, or often, at best scraping of web pages for information. While this is not the right approach, it is often the only available approach. Standardized service interfaces are needed for interfacing data with modeling and visualization tools. Data formats should be standardized through community use cases. Data product needs for earthquake science are as follows:

* All data products should be coupled with self-describing network service interfaces. A great deal of useful data and metadata about earthquakes, for example, is bound in human-readable web pages instead of machine-readable formats (e.g., ontologies).
* Services should be documented, published, and discoverable.
* Services for analyzing lower level data products should also be designed with the same approach. These services generate products that may be consumed downstream.

Data presented in a map view that can be browsed eases selection of the data (Figure 2). Information about data is often encoded into long file names, and often locations over which the data are collected are encoded as station names or flight paths. Without the user having familiarity with the identification scheme it is often difficult to locate data of interest. A map view of the data makes it easier for a user to efficiently scan for data over regions of interest. Problems arise when data are collected over different time spans and overly other data in the same region, or multiple data interpretations exist. Pop-up lists, menus, or time slider bars can alleviate some of these issues. The data in the QuakeTables database are also can be accessed through APIs, which connects the data directly to various applications.

4. Computational Infrastructure

A user friendly computational infrastructure is necessary for identifying and pulling in data from numerous sources, simplifying or automating data assimilation, mining, and modeling workflow, and providing feeds and interfaces for generalized data users. QuakeSim provides a back-end for earthquake forecasting and response, crustal deformation modeling, and modeling of fluids within the crust. The scaling of compute power should occur on the back end and be transparent to the user.

QuakeSim applications require the user to do the following either in an automated manner or with user intervention: 1) Select data in terms of types, time, and space; 2) Subset data to relevant focus of interest; 3) Move data for mining, modeling, or visualization; 4) Analyze data by modeling, inverting, or data mining; 5) Visualize data and results; and 6) Track data and models. For small data sets or regions of interest these steps can be done manually and in fact such investigations provide excellent examples for developing workflow for larger and more complicated cases. Current data volumes and in particular those for existing or planned InSAR missions motivate the need for an end-to-end architecture in which data can be systematically analyzed, modeled, and interpreted. Automation requires interfaces between the widely distributed data sets, data products, and applications. Without such a system in place at launch the vast majority of the planned DESDynI-R mission data will be under or not utilized.



Figure 3. QuakeSim portal showing anonymous crustal deformation model tool with fault map browser.

In an end-to-end computational infrastructure users should be able to evaluate data, develop science models, produce improved earthquake forecasts, and respond to disasters in intuitive map-based interfaces. Fault models can be constrained and improved not just by geology, but also by feature identification from InSAR (UAVSAR) and inversions of both GPS and InSAR crustal deformation data [2]. Forecasting is improved by development of better interacting fault models, pattern analysis, and fusion of both seismicity and crustal deformation data. Increasing the accessibility and utility of GPS, InSAR, and geologic data, addresses science challenges such as earthquake forecasting or fluid migration. Intuitive computational infrastructure also can enable new observations by providing tools to conduct simulation experiments and new information products for use in a wide variety of fields ranging from earthquake research to earthquake response. Timely and affordable delivery of information to users in the form of high-level products is necessary for earthquake forecasting and emergency response, but it also necessary for exploiting crustal deformation to enable new discoveries and uses.

There are numerous practical issues to establishing an effective computational infrastructure. Of chief importance is that tools be intuitive and easily accessible. Some QuakeSim tools are public and reside outside of any required login. This mode of operation is often preferred by users as it avoids the need to remember another login and password combination and allows for greater privacy. However, there are also limitations. Chief of these is that project tracking is not possible. The user would be required to maintain projects locally, which is reasonably easy with simple input and output files, but becomes rapidly complicated when project components are coupled to various applications at the back end. For example, a user may set up and run a model, which is then coupled to various output format and map views. Linking project components is easier if it is done at the back end within a logged in environment. It also is more efficient in an environment where large data sets are accessed and/or displayed.

Data-intensive computing infrastructure provides a modeling and visualization environment to a broad geophysical community, supporting multiple data types without the need to download large data sets. Access to GPS, InSAR, faults models, and seismicity is just starting to be coordinated: today, large amounts of data must move to the investigator's computer, and integration into models is ad hoc. Modeling interacting-fault simulations largely takes place on local efforts at the research group level, with comparisons taking place largely at infrequent workshops. Web-service based interfaces allow public, independent verification and comparison of simulators and statistical forecast methods, feeding directly into regional hazard models.

Data discovery is also ad hoc and likely to miss important elements of data fusion and cross validation. Ontology-based methods allow immediate discovery of topically or space-time related data. Rather than bringing all data to the user, a system will increasingly need to place substantial processing in cloud computing services close to original or mirrored data archives. Designing around systems from data to investigators will encourage widespread use of enormous data collections such as gathered by radar missions, rather than ad hoc use by a small community of experts. It will be increasingly necessary to couple computing capabilities to the data storage using cloud computing approaches for management of very large data sets.

The computational requirements of many current QuakeSim runs are modest but QuakeSim is architected so that it can scale up by the many orders of magnitude needed when new satellites or other instruments result in much larger data sets. Iterative MapReduce [5-8] that interoperates between HPC and cloud environments will be deployed to handle the much larger datasets and model runs.

5. Visualization

Visualization is often necessary for interpreting data or models. Challenges exist both in visualizing complex data as well as in producing visualizations that are properly constrained by data. Visualizations should also be flexible so that the user can view the data or model output in different ways. For example, GPS velocities vectors, when plotted relative to different stations illuminate different features responsible for the deformation (Figure 4). When GPS vectors are plotted relative to the San Gabriel Mountains compression to the west in the Ventura basin becomes clearly apparent. Shear zones on either side of the San Gabriel Mountains are also apparent. Similar plots, but relative to stations in the Mojave Desert highlight the Eastern California Shear Zone.

Movies of deformation can show transient as well as secular deformation. Visualizations are useful to when very small changes in the actual observation data are exaggerated in order to be visible to the viewer, and to show a very long timeframe changes displayed in a compressed time, as long as the exaggeration ratio is chosen properly. However, challenges arise with accurately driving the animation from the data. GPS stations are sparsely located as indicated in Figure 4 and as a result interpolations between the stations must be made. Additionally, GPS time series do not all exist for exactly the same time frame, which introduces meshing complexity or adds artifacts to the visualization from station outages. GPS time series must be properly interpolated both spatially and temporally to provide the most physically accurate animation. InSAR data are also sparse and occur typically for short timespans, but can further guide mapping of crustal deformation. UAVSAR observations in southern California have identified numerous localized zones of shear that could not be identified with the spatial sampling provided by GPS.

6. Interfacing Applications

It is not likely, nor is it necessarily desirable, for one monolithic computational infrastructure to be developed for accessing and modeling remotely sensed geodetic data. Centers of expertise are distributed geographically and these centers are providers of components of data, data products, models, or information that lie upstream or downstream from other components. For example, fault modeling connects observational data sets to downstream simulation and forecasting techniques. Ideally these various centers of expertise will develop computational infrastructure that interfaces with upstream or downstream data, data products, applications, or information. Such a system does not currently exist to support solid earth research, but steps are being taken in that direction as groups realize the need to interface with either upstream or downstream components.



Figure 4. GPS vectors plotted for southern California relative to station CHIL in the San Gabriel Mountains. Fault traces are from UCERF-2 [4].

NASA’s Distributed Active Archive Centers (DAACs, http://nasadaacs.eos.nasa.gov/) process, archive, document, and distribute data from NASA’s various missions and field activities. The DAACs serve one or more specific Earth science disciplines and provides data products, data information, services, and tools unique to its particular science. Historically analysis was done close to the DAACs. However, as tools and analysis groups become more distributed it is important to ensure that where necessary data are mirrored or are otherwise close to the applications to avoid lengthy transfers of large data sets to accomplish tasks.

7. Conclusions

Increasing data volumes, complexity of data processing algorithms, and more comprehensive models continually drive a need for more and more compute power. Additionally, as large data sets are accessed it is necessary to either keep the data close to the models or have extremely broadband connections between the data sources and computers carrying out the models. We have suggested that data repositories co-located with cloud computing is an attractive approach.

A critical requirement is that the environment support both science and disaster community users.

Large jobs are currently run on supercomputers, which reside at high performance computing facilities. These resources are often oversubscribed and users’ models can spend a long time in a queue before the job runs. Local machines may not be adequate for running large models, however. Investment needs to be made in more high performance computers and facilities or investment needs to be made in an elastic cloud-computing infrastructure that has high communication bandwidth between nodes.

Visualization tools are increasingly necessary for understanding data and models. Users are frequently hampered by tools that rely on licensed products or not run in the same environment in which the data or models are stored or processed. This results in the need to move large volumes of information and often requires an additional reprocessing or reformatting step before visualization can take place. Open source tools are not yet mature enough. They are sometimes not maintained or are incomplete. An investment in open source visualization tools will result in much greater scientific efficiency. Both data management and simulation tools would benefit from a redesign of the underlying computing infrastructure.

Acknowledgements

This work was carried out at the Jet Propulsion Laboratory, California Institute of Technology, Indiana University, University of Southern California, and University of California’s Davis and Irvine campuses under contract with NASA. The work was sponsored by NASA’s Advanced Information Technologies Program. We thank UNAVCO, SOPAC, ASF, and University of Nevada, Reno for ongoing collaborations.

References

1. Donnellan, A.; Rosen, P.; Graf, J.; Loverro, A.; Freeman, A.; Treuhaft, R.; Oberto, R.; Simard, M.; Rignot, E.; Kwok, R.; Xiaoqing Pi; Blair, J.B.; Abdalati, W.; Ranson, J.; Zebker, H.; Hager, B.; Shugart, H.; Fahnestock, M.; Dubayah, R.; , "Deformation, Ecosystem Structure, and Dynamics of Ice (DESDynI)," *Aerospace Conference, 2008 IEEE* , vol., no., pp.1-13, 1-8 March 2008  
   doi: 10.1109/AERO.2008.4526249
2. Donnellan, A., Parker, J.W. and Peltzer, G., Combined GPS and InSAR models of postseismic deformation from the Northridge earthquake, *Pure Appl. Geophys.*, **159**, 2261-2270 (2002)
3. Wei, M, D Sandwell and B Smith-Konter, Optimal combination of InSAR and GPS for measuring interseismic crustal deformation, *Adv. Space Res*., **46**, 236-249 (2010).
4. Field, E.H., T.E. Dawson, K.R. Felzer, A.D. Frankel, V. Gupta, T.H. Jordan, T. Parsons, M.D. Petersen, R.S. Stein, R.J. Weldon II, and C.J. Wills, The Uniform California Earthquake Rupture Forecast, Version 2 (UCERF 2), By 2007 Working Group on California Earthquake Probabilities, USGS Open File Report 2007-1437, CGS Special Report 203, SCEC Contribution #1138, 2008.
5. SALSA Group. Iterative MapReduce. 2010 [accessed 2010 November 7]; Twister Home Page Available from: http://www.iterativemapreduce.org/.
6. J.Ekanayake, H.Li, B.Zhang, T.Gunarathne, S.Bae, J.Qiu, and G.Fox, Twister: A Runtime for iterative MapReduce, in Proceedings of the First International Workshop on MapReduce and its Applications of ACM HPDC 2010 conference June 20-25, 2010. 2010, ACM. Chicago, Illinois. http://grids.ucs.indiana.edu/ptliupages/publications/hpdc-camera-ready-submission.pdf.
7. Bingjing Zhang, Yang Ruan, Tak-Lon Wu, Judy Qiu, Adam Hughes, and Geoffrey Fox, Applying Twister to Scientific Applications, in CloudCom 2010. November 30-December 3, 2010. IUPUI Conference Center Indianapolis. http://grids.ucs.indiana.edu/ptliupages/publications/PID1510523.pdf.
8. Thilina Gunarathne, Judy Qiu, and Geoffrey Fox, Iterative MapReduce for Azure Cloud, in CCA11 Cloud Computing and Its Applications. April 12-13, 2011. Argonne National Laboratory, Chicago, ILL. http://grids.ucs.indiana.edu/ptliupages/publications/cca\_v8.pdf.