Integrating Remotely Sensed and Ground Observations for Modeling, Analysis, and Decision Support

Andrea Donnellan, Margaret Glasscoe, Jay
W. Parker, Robert Granat
Jet Propulsion Laboratory
California Institute of Technology
4800 Oak Grove Dr.
Pasadena, CA 91109
818-354-4737
Andrea.Donnellan@jpl.nasa.gov
Margaret.T.Glasscoe@jpl.nasa.gov
Jay.W.Parker@jpl.nasa.gov
Robert.Granat@jpl.nasa.gov

Marlon Pierce, Jun Wang, Geoffrey Fox
Indiana University
2719 East 10th Street
Bloomington IN 47408
marpierc@iu.edu
wang208@indiana.edu
gcf@indiana.edu

Dennis McLeod
University of Southern California
Computer Science Department
Los Angeles, CA 90089-0781
mcleod@usc.edu

John Rundle, Eric Heien
University of California, Davis
One Shields Avenue
Davis, CA, 95616
jbrundle@ucdavis.edu
emheien@ucdavis.edu

Lisa Grant Ludwig
University of California, Irvine
653 E. Peltason, AIRB 2085
Irvine, CA 92697-3957
lgrant@uci.edu

Abstract—Earthquake science and emergency response require integration of many data types and models that cover a broad range of scales in time and space. Timely and efficient earthquake analysis and response require automated processes and a system in which the interfaces between models and applications are established and well defined. Geodetic imaging data provide observations of crustal deformation from which strain accumulation and release associated with earthquakes can be inferred. Data products are growing and tend to be either relatively large in size, on the order of 1 GB per image with hundreds or thousands of images, or high data rate, such as from 1 second GPS solutions. The products can be computationally intensive to manipulate, analyze, or model, and are unwieldy to transfer across wide area networks. Required computing resources can be large, even for a few users, and can spike when new data are made available or when an earthquake occurs. A cloud computing environment is the natural extension for some components of QuakeSim as an increasing number of data products and model applications become available to users. Storing the data near the model applications improves performance for the user.

Table of Contents

1. INTRODUCTION

QuakeSim is a computational environment integrating data and models for understanding earthquake processes. QuakeSim focuses on using remotely sensed geodetic imaging data to study the interseismic part of the earthquake cycle (Figure 1). Geodetic imaging data provide observations of crustal deformation from which strain accumulation and release associated with earthquakes can be inferred. The goal is to understand earthquakes for mitigation and response. Remotely sensed geodetic imaging observations provide information on motions of the surface of the Earth's crust, expressing the cycle of strain accumulation and release. These data products are fused with geologic and seismicity information in the QuakeSim environment with visualization, modeling, and pattern analysis tools. E-DECIDER is an earthquake response decision support system that also uses remotely sensed observations and some QuakeSim components.

Earthquakes are the result of complex solid Earth processes. Plate tectonics drive motion of the Earth's crust, accumulating deformation primarily along plate boundaries. Strain stored as elastic energy is released in earthquakes along faults. These processes occur on scales of milliseconds to millions of years, and from microns to global scales. Crustal deformation data products derived from GPS and interferometric synthetic aperture radar (InSAR) observations enable an understanding of the cycle of strain accumulation and release. Geological and seismic data complement the spaceborne observations. Various domain experts study aspects of these processes using experimental, theoretical, modeling, or statistical tools. Over the last decade, in part due to the advent of new
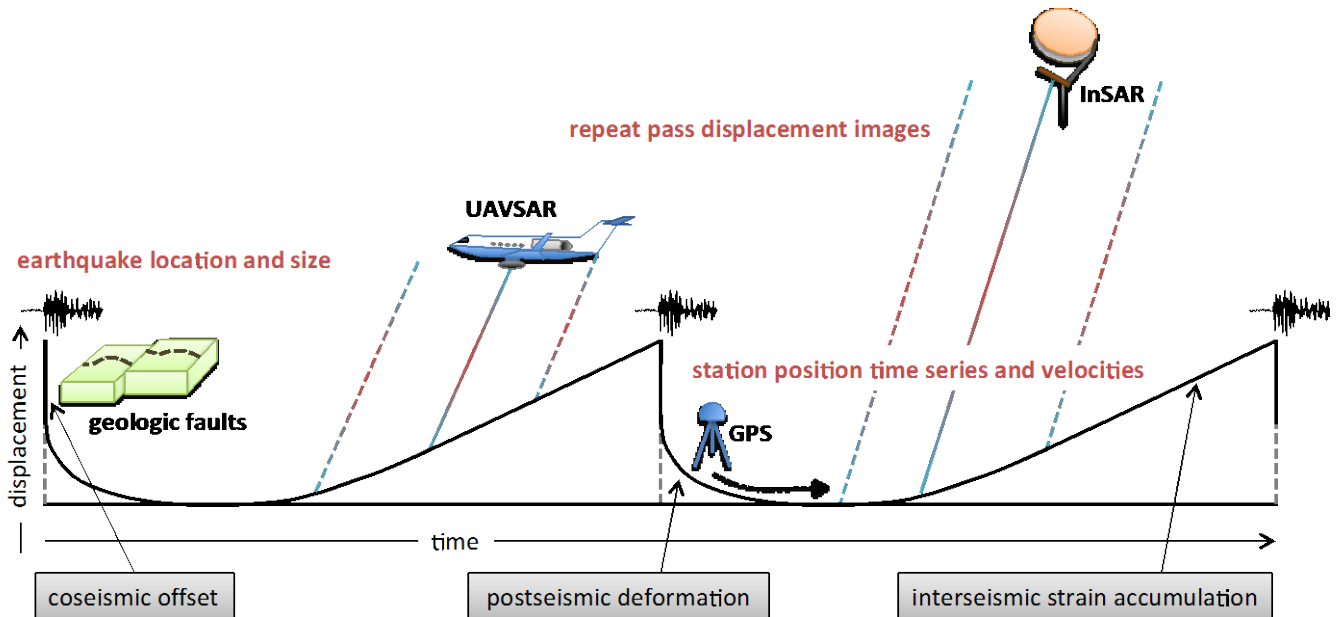
Figure 1. Schematic of the earthquake cycle of strain accumulation and release with the long-term tectonic rate removed.

computational methods, there has been an increasing focus on mining and integrating the heterogeneous data products into complex models of solid Earth processes.

QuakeSim focuses on the cycle of strain accumulation and release that is well addressed by geodetic imaging data and related model applications. Coseismic offset is the slip that occurs during an earthquake. Postseismic deformation is continued motion that occurs following an earthquake and deviates from the long-term slip rate. Postseismic deformation is usually some combination of afterslip, relaxation, or poroelastic rebound. Interseismic strain accumulation occurs as a result of plate tectonic motion. QuakeSim addresses the size and location of events, coseismic offsets associated with the earthquakes, and the time varying displacements that occur throughout the earthquake cycle. Seismicity and geology yield information on size and location of events. GPS provides position time series. Interferometric synthetic aperture radar from airborne and spaceborne platforms provide images of displacement. All data sources provide information on geometry of faults and time varying slip.

QuakeSim and E-DECIDER are downstream data product consumers that perform forward modeling, inverse modeling, forecasting, event detection, and response on the data products. Both E-DECIDER and QuakeSim rely on processed data products, which are typically velocities and time series from GPS, and ground range changes from Interferometric Synthetic Aperture Radar (InSAR). These and other data products such as seismicity, geologic fault data, or civil infrastructure, are integrated and assimilated into models for understanding, forecasting, and response. As a result of this reliance on data product providers, it is imperative that data specifications, standards, and locations are well understood.

QuakeSim is distributed and heterogeneous because of the many different data products and numerous different applications; it is consequently implemented using Service Oriented Architecture principles. Separating QuakeSim into its different parts, strategically using cloud resources where necessary, and using other resources as practical result in greater efficiency. In cases where data volumes are large, such as InSAR, it is important to ensure that model applications reside close to the data products to reduce latency from bandwidth limitations. Alternatively, data products can be connected to applications through high bandwidth pathways. Services such as Amazon Glacier are not adequate because they have a long latency of 3-5 hours for retrieval of data. Scientists need to be able to browse data products and analyze them over short time periods, particularly for event response.

While we use seismicity as a data source for size, location, and mechanism of events, we do not focus on seismic waveforms. Understanding crustal deformation and fault behavior leads to improved forecasting, emergency planning, and disaster response. Analysis of crustal deformation data can be used to indicate the existence of otherwise unknown faults [1], particularly when they are not exposed at the surface. Accurate fault models are required for understanding earthquake processes and require the integration of multiple types of data. Identifying, characterizing, modeling and considering the consequences of unknown faults, and improving the models of known faults contributes to seismic risk mitigation.
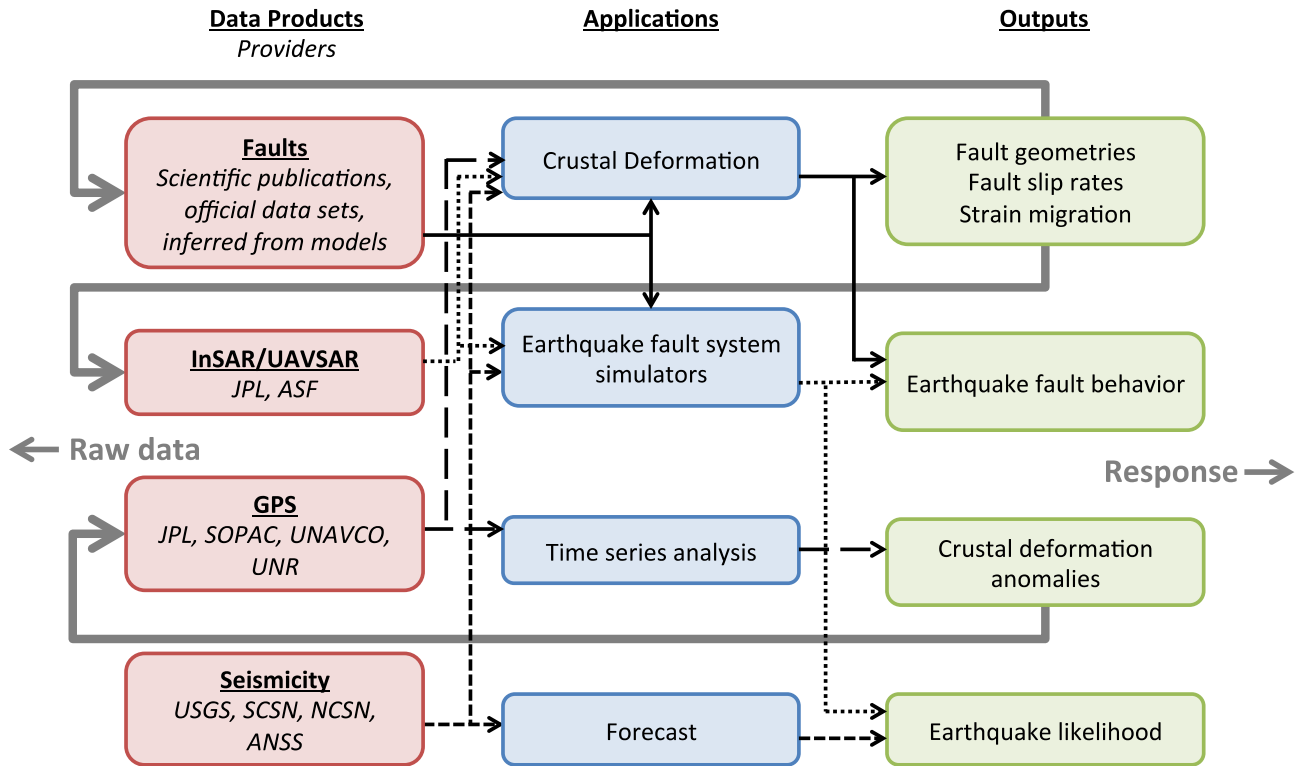
Figure 2. QuakeSim inputs, outputs, and feedback to data providers are shown. Multiple data products are used from numerous providers. These are used in applications for understanding faults and forecasting earthquake likelihood. Output from the applications often identifies issues with the data products, which can be in the form of previously unrecognized faults, unmodeled error sources, or processing issues.

The goal of QuakeSim is to improve understanding of earthquake processes by integrating remotely sensed geodetic imaging and other data with modeling, simulation, and pattern analysis applications. Model and analysis applications are delivered as standalone offline software packages or online through web services and Science Gateway user interfaces [2]. Data products are available online for browsing, download, and online or offline analysis through a database and services. QuakeSim utilizes data products that have typically been generated elsewhere; raw data is not ingested or processed but higher level data products from other data providers are made available through Open Geospatial Consortium (OGC)-compliant web services [3,4].

2. DATA PRODUCTS

QuakeSim applications use multiple data products from multiple providers (Figure 2). Data products include faults, GPS position time series and velocities, Radar repeat pass interferometry (RPI) or Interferometric Synthetic Aperture Radar (InSAR) ground range change images, and seismicity. It is important that the data products come from trusted sources and are reliable in order to construct valid models of earthquake processes and produce accurate earthquake probabilities. Understanding the history of the data products and how they were produced allows a user to make informed decisions and interpretations. It is also important to ensuring reproducibility of results. Tracking data provenance is necessary, particularly when data are reprocessed or originate from a variety of solutions, such as with GPS data. For individual derived data products and integrated data products the tracking and appropriate propagation of accumulated error and precision is particularly important.

The requirements for QuakeSim as a data product consumer and upstream users of QuakeSim results are the same. Standards and services are necessary for efficient consumption and analysis of products. Data product customers specifically need: 1) Standard data formats like GeoTIFF (UAVSAR), GeoJSON, etc.; 2) Better metadata, such as how the data product was explicitly created (particularly for GPS, in our experience); 3) No undocumented, post-publication changes; 4) Delivery of data via standard mechanisms, such as via OGC-based Web services; 5) Cloud-based delivery of data sets, such as with Amazon S3 and Elastic Block Store services; and 6) Notification services for new products and changes. Web Services are key to effectively supporting downstream users; these should not be entangled with Web and desktop clients. Effective utilization of data products and analysis results by emergency responders require standard interfaces and products that are easily and readily interpretable.

Inputs   Derived Products   Delivery

Remote Sensing Imagery

QuakeSim Tools

Optical Change Detection

Deformation Change Detection (Simulations, GPS, InSAR)

HAZUS Input Generation & Critical Infrastructure Information

Earthquake & Aftershock Forecasts

UICDS Data Sharing Middleware

GIS compliant map products and KML output
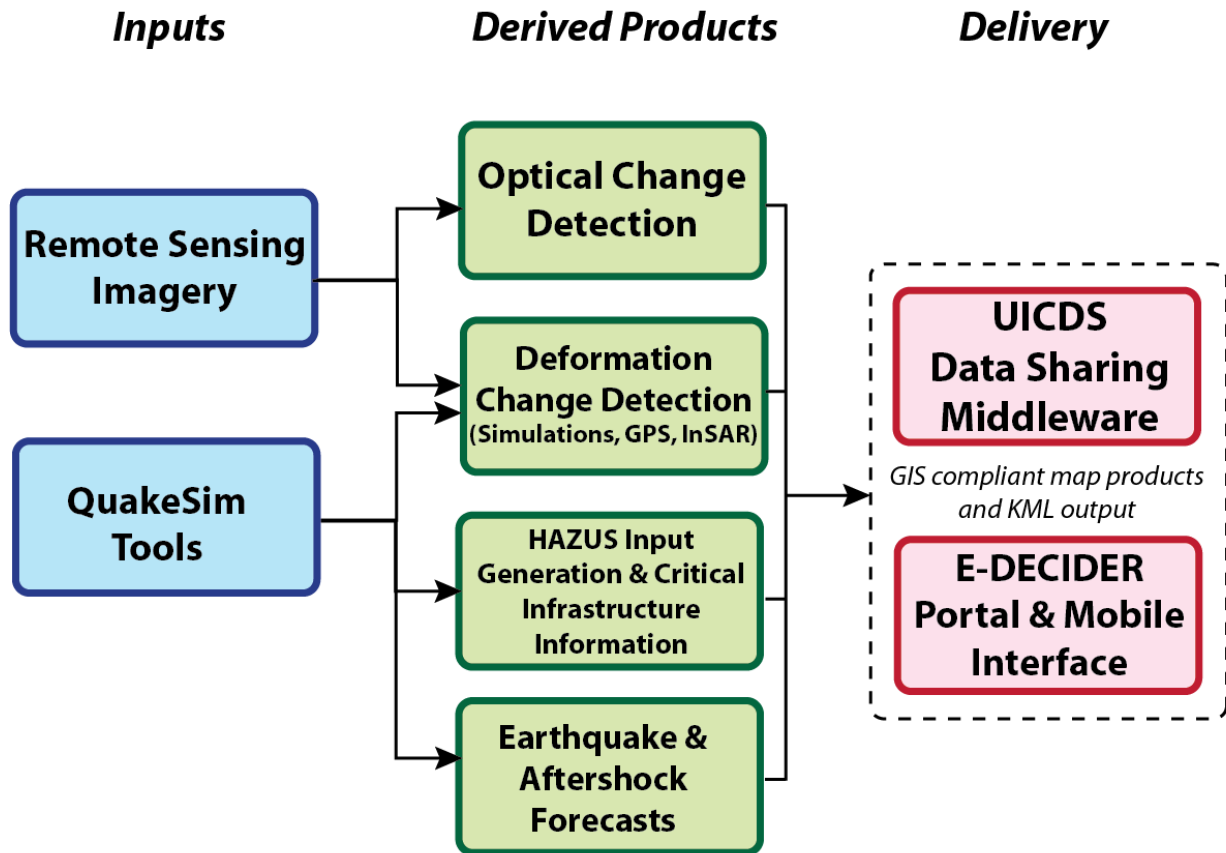
E-DECIDER Portal & Mobile Interface

Figure 3. Simplified E-DECIDER workflow where inputs, outputs, and delivery mechanisms are shown. Input data in the form of remote sensing imagery or simulation results generate derived decision support products that are then delivered as OGC-compliant products through the DHS UICDS software and the E-DECIDER interfaces.

A key issue for earthquake response efforts is data availability and reliability, especially in the first hours to days following the event. Responders and decision makers require access to rapid fault solutions, deformation and displacement information, and geologic and geodetic data quickly and easily. Available and trusted data products in times of crisis are essential in the decision making process. Quick-look products that may not be as accurate are of more value than accurate products with a long latency.

3. MODELING AND ANALYSIS

Data products are used in four main ways by QuakeSim modeling and analysis tools. The goal is to identify active faults or regions, understand earthquake fault behavior, search for anomalies, and estimate earthquake likelihood. QuakeSim provides integrated map-based interfaces and applications for 1) accessing remotely-sensed and ground-based data products; 2) exploring and analyzing observations; 3) modeling earthquake processes; and 4) conducting pattern analysis to focus attention and identify significant and/or subtle features in the data. Data products and model tools can be accessed by network service (Web Services and REST) interfaces, directly through web-based user interfaces, or run offline on local or high-performance computers.

Cloud computing is applicable in cases where multiple runs can be made in an embarrassingly parallel fashion. This includes analysis of high rate GPS time series analysis, during post event response, when numerous users might be using the QuakeSim environment, and for simulating earthquakes or inverting crustal deformation observations in which numerous runs are carried out with different initial conditions.
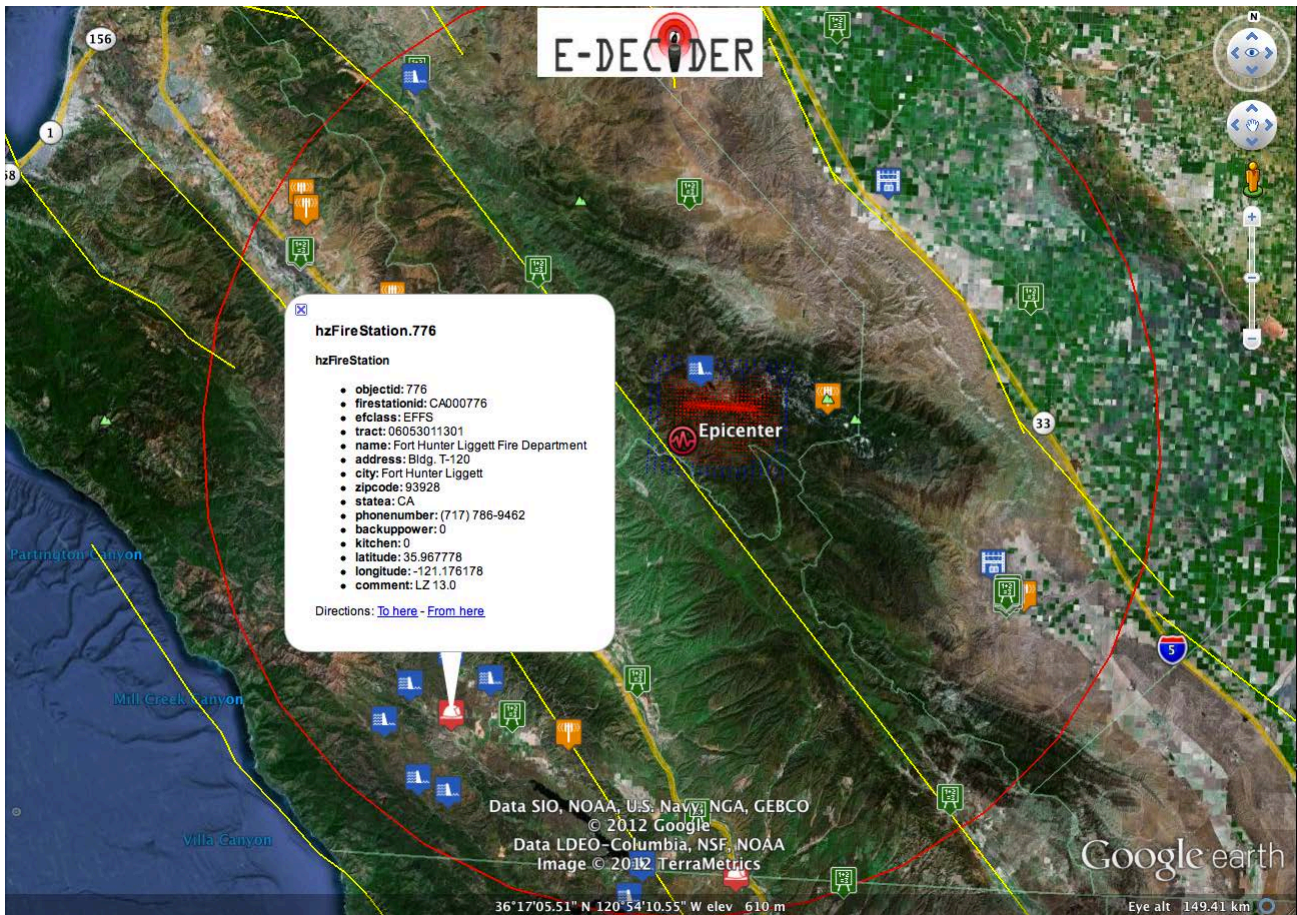
Figure 4. Example E-DECIDER output with deformation simulation result (tilt map – red/blue dots) generated from QuakeSim RSSDisloc tool with epicenter and 40 mile radius denoted with E-DECIDER's On-Demand HAZUS KML Generation service. Critical infrastructure can be overlain and attention can be focused on potential facilities that may have been exposed to damage resulting earthquakes. The call out bubble identifies a fire station within the 40 mile radius of the epicenter of the 21 October M 5.3 Central California earthquake.

Data products and modeling results are treated in a framework that includes per-sample uncertainties and propagation of errors. This framework is not yet complete, because some data products (particularly InSAR phase) do not come with a full description of correlated errors, and because some forward modeling applications (such as viscoelastic models) do not support direct propagation of errors. However, the current treatment of uncertainty does include propagation of errors, such as propagation of GPS time series per-sample uncertainty (reported from a prior positioning estimation application) through an inversion application that results in a model of fault geometry and slip that includes the propagated errors. Another advantage to inversion with errors is that inadequate sensor geometry and accuracy may be detected by large propagated errors. In addition the generation of a suite of forward models based on fault geometry and slip conditions within a defined uncertainty space will generate a suite of synthetic surface displacements that propagate the characteristics of that initial uncertainty space. Investigations of this latter sort are well suited to cloud computing due to their uncoupled parallel nature.

4. DECISION SUPPORT

QuakeSim tools and Web services are used as a back end for E-DECIDER. E-DECIDER provides decision support for earthquake disaster management and response utilizing NASA and other available remote sensing data and QuakeSim modeling software (Figure 3). E-DECIDER delivers a web-based infrastructure and mobile interface designed for ease-of-use by decision makers, including rapid and readily accessible satellite imagery and synthetic interferograms following earthquakes; standards-compliant map data products; and deformation modeling and earthquake aftershock forecasting results [4]. The decision support tools are developed working in partnership with end users in first response and disaster management agencies, such as the US Geological Survey (USGS), California Geological Survey (CGS), and the Department of Homeland Security (DHS). A recent addition to the suite of tools includes delivery of E-DECIDER information through the DHS Unified Incident Command and Decision Support (UICDS) software, which allows data providers a means to

communicate and send their products in a standard manner in the event of a disaster.

An example workflow includes an end-to-end process that starts with the triggering of a deformation simulation from a >M5 earthquake from the USGS feed and ends with those results posted to an RSS feed. We have developed algorithms and web services to automatically process output from the Disloc model to produce tilt and deformation gradient information using a Sobel operator. The tilt is a measurement of the change of slope that can affect water distribution, drainage, and sewage services. The deformation gradient is a measure of the rate of change of deformation and can detect locations of possible fault rupture on the surface as well as locations where infrastructure may be damaged by tension, compression, or shear at the surface.

The tilt and gradient processing can be performed as an on-demand process or as a web service. The web service version will be chained with RSSDisloc to provide automatic mapping of tilt and gradient for earthquakes as they appear on the USGS Prompt Assessment of Global Earthquakes for Response (PAGER) feed. The tilt and gradient information can be mapped either with KML for display in Google Earth or with ArcGIS or other desktop GIS tools. Mapping products generated in ArcGIS are shared either as image/PDF files or in native mapping formats. Additionally, these products will be disseminated automatically via UICDS to responders (in locations where UICDS is in use). The JPL UICDS core has been installed and is operational and has been tested in a response exercise to push E-DECIDER products to end users in an operational environment during the ShakeOut in mid-October 2012. Future plans to send epicenter and potentially exposed infrastructure from Hazards-United States (HAZUS) databases from the E-DECIDER GeoServer directly to UICDS upon the USGS feed trigger are also underway.

5. CLOUD COMPUTING

Cloud computing in the general sense represents the centralization of computing capacity into highly scalable data centers [5]. The National Institute of Standards is standardizing cloud computing concepts and terminology, which we briefly summarize here.

In brief, we can think of clouds as consisting of the following service layers. *Infrastructure as a Service:* through Virtual Machine technologies, clouds provide access to hosted servers and resources that appear to the user as "just another server." The advantage from the user's point of view is that the resources are elastic and can be scaled up or down as needed. Amazon, Azure, and Google Compute Engine are examples. Open source examples include OpenStack, OpenNebula, CloudStack, Nimbus, and Eucalyptus. *Platform as a Service:* This allows the user to deploy user-created applications within a specific cloud framework into the cloud for use by others. Google App Engine and Amazon's Elastic MapReduce [6] are prominent examples. Here the cloud provider creates a framework that can be used to build applications while hiding underlying complexities (such as those exposed in Infrastructure as a Service). *Software as a Service:* The cloud provider provides a specific capability or software to the user. Examples are seen in the individual services corresponding to different QuakeSim tools or to individual services corresponding to different forecasting models.

Compared to traditional high performance computing, Cloud Computing provides potentially greater computing capacity but with larger latencies. This makes Cloud Computing an excellent candidate for "big data" problems in which the same operations need to be performed on very large data sets, but little communication between individual processes or movement of data is needed. Further, even in problems needing communication, these are not small messages familiar from parallel simulations but rather large "reductions" (such as global sums or broadcasts) supported by MapReduce and its iterative extensions in the cloud [7-10].

A growing body of research indicates that Cloud computing approaches are a good match for many large-scale data-centric scientific computing problems [10-16]. Cloud computing infrastructure in many fields is overtaking the traditional data center. The size of data in many fields (such as the life sciences) is growing so rapidly that frequent data movement is impractical, and so computing must be brought to the data. This is likely the case for InSAR geodetic imaging data as volumes grow over time. High rate GPS data products also are good candidates for analysis on the cloud. Other applications may be impractical for cloud computing, and analysis should be done to gain efficiency with cloud computing while not incurring expense where unnecessary.

It is important to prototype these approaches for NASA data collections, particularly the anticipated large collections of InSAR imagery from the DESDynI-R mission. At its most basic level, Infrastructure as a Service allows us to quickly experiment with different resources (number of CPUs, memory size) to determine the optimal resource requirements for running a service. However, one can obtain similar flexibility from modern, Virtual Machine-based data centers such as Indiana University's Intelligent Infrastructure service. We need to investigate the potential advantages of Cloud Computing over traditional data centers. As described above, there are three possible Cloud Computing scenarios within QuakeSim.

First, QuakeSim Web services, such as OGC-based services for delivering UAVSAR data, can be moved to clouds from their current locations in university data centers. The advantage of clouds such as Amazon over traditional data centers for QuakeSim's stateless services include *auto-scaling, load balancing,* and *content distribution.* Auto-scaling allows us to automatically scale up or down the

number of server instances used by a given service. This can be done using various triggers, such as the detection of high usage or on regular schedules. This is a key feature of elasticity in clouds. Load balancing is the distribution of a service across multiple services with typically round-robin distribution of requests. This also provides greater fault tolerance if a particular service instance is down. Although this can be configured manually in traditional Apache HTTPD-based services, many cloud providers build load balancing into their fabric, requiring no modification of the service. Finally, content distribution allows the services to be replicated geographically and located on network edge servers, potentially reducing the latency of requests. This is particularly interesting to investigate for QuakeSim, given the world-wide interest in earthquake events and our requirements for building interactive tools.

Second, cloud computing "platform as a service" offerings such as map-reduce and its variations are very interesting models for data intensive computing with several recent earth/environmental science examples including GIS services in the cloud [17-23]. Potential applications include applying feature detection algorithms to InSAR and UAVSAR data sets to potentially identify previously unknown faults as well as the Virtual California toll described later. Recent "high performance computing" cluster resources can also be obtained from Amazon. These are more traditional clusters, exchanging virtualization for higher speed networking. These provide an alternative to traditional campus clusters that is appropriate for more traditional high performance computing [12, 16].

Third, cloud computing provides a way for QuakeSim to work with its collaborators to more efficiently share data sets (as in Figure 2). As described previously, QuakeSim does not process raw observational data but is instead downstream consumer of data. Maintaining consistent copies with our upstream data providers is a challenge. QuakeSim's QuakeTables database houses some processed InSAR data products and also the complete set of processed repeat pass interferometry products from the airborne UAVSAR InSAR project. These data provide essential information for modeling earthquake processes and particularly for developing accurate fault models. We are collaborating with data product providers to ensure standard interfaces formats as well as jointly used cloud infrastructure where appropriate. The infrastructure must be flexible enough to support other data sets and use cases.

Under the present cloud models, storage at existing data center appears more cost effective than storage on the cloud where recurring costs are at present prohibitive. However, this may change in the future. Microsoft Azure's Blob storage service, Amazon's S3, and the Lustre file system-based Whamcloud are examples of unstructured storage, and BigTable, HBase, and the Azure Table Service are examples of structured data storage. We will evaluate these for the storage and access of large collections of individually large data sets. A key observation from our

research on cloud systems for science is that data storage and computing must be coupled, making systems such as Amazon Glacier, with slow retrieval times inadequate. We must therefore couple our IaaS prototyping with SaaS prototyping. The SaaS models that we are considering are MapReduce and its derivatives. MapReduce style approaches are particularly useful for data-parallel computing problems such as L-band radar image processing and GPS signal analysis. Simulations and inversions where similar runs are done with different initial conditions are also candidates for cloud computing.

We have carried out some initial testing on different platforms. WCS (Web Coverage Service) is an OGC standard to retrieve raster data based on spatial constraints and other query criteria. For our testing procedure we 1) Randomly choose one InSAR image, 2) Randomly generate a small rectangle bounding box inside an InSAR image, and 3) Send a request to GeoServer and retrieve the image in GeoTiff format. Performance is measured by the average response time of one user request call, which includes GeoServer response time and network transfer time.

The performance WCS is tested with three different requests with the same bounding area: first is to retrieve image in native resolution, the second requests the image in lower resolution by resampling the original image, and the third one requests the image in lower resolution and different projection. The tests are carried out on GeoServers running on Indiana University Intelligent Infrastructure (IUII) and Amazon EC2 Cloud. Table 1 shows the results of clients from both IU and Amazon. It shows there is no substantial performance difference between cloud service and traditional data center for this simple test case. More realistic testing are focused load balancing and scalability to simulate the spikes of data and user requests during earthquake events.

Table 1. WCS performance in seconds

	Test Client from IU		Test Client from Amazon	
	IUII	Amazon	IUII	Amazon
Native Resolution	0.57	1.28	1.30	0.54
Resample	0.27	0.73	0.70	0.23
Resample + ReProjection	0.34	0.78	0.71	0.24

Cloud computing systems such as FutureGrid [24] also provide infrastructure for high performance earthquake simulation. This system can be dynamically reconfigured between virtual machines and bare-metal allowing both high performance and measurement of virtualization overhead. Figure 5 shows parallel scaling and performance results from the QuakeSim Virtual California earthquake simulation code on the FutureGrid system. These tests were run using a 3000 year simulation of seismicity in a model

containing all major faults in California.

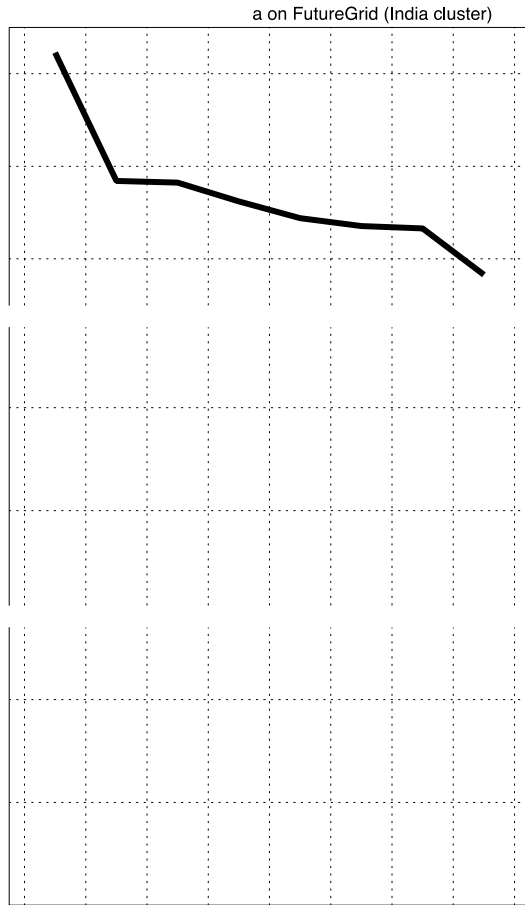The results show that VC scales well up to 64 cores (8

a on FutureGrid (India cluster)

Figure 5. Parallel scaling and performance analysis of earthquake simulator Virtual California on FutureGrid test bed.

nodes) on such a system at which point communication dominates and takes over half of runtime. The bottom figure shows the fraction of runtime spent in different sections of the program, with calculation represented by shades of blue and communication by shades of red. The middle figure shows the total wall time of the simulation, including the wall time for calculation and ideal scaling. The top figure shows the memory bandwidth achieved during the matrix-vector multiply which is the dominant calculation during the simulation and is generally limited by bandwidth due to the ratio of bytes to flops.

There are several points related to scalability. First, as shown in the lower figure, communication as a fraction of total runtime stays relatively constant from 4 cores up to 32 cores, and significantly grows past 64 cores. This is likely due to the Infiniband network configuration where up to 8 nodes may be connected on a single switch. Past 8 nodes communication must travel between switches, which accounts for the jump in communication time. Each type of

communication is global in nature so their scaling is similar.

Another issue is related to scaling of the calculation. Since the matrix in the matrix-vector multiply is dense, the limiting factor is generally memory bandwidth. With 2 cores a large fraction of the full memory bandwidth (10.4 GB/s out of a theoretical 12.4 GB/s) can be utilized. However, adding more cores only serves to divide this limited bandwidth further, dropping per-core bandwidth to 7.6 GB/s (this is higher than the theoretical limit because of on-core cache) and lower as the number of cores increases.

Based on these results, further calculation performance improvements in VC must focus either on compression or different algorithmic approaches to reduce memory bandwidth, or utilization of fewer cores on more nodes to better distribute the bandwidth limitation. Techniques to improve communication performance in larger clusters are also applicable such as hierarchical communication within then between nodes.

6. SUMMARY

Clouds are a promising new method for analysis and modeling of large data product sets and for multiple parallel runs. In some instances it may be better to use standard resources instead of clouds. Careful consideration should be given to what resources to use for various tasks. QuakeSim consists of several components interfacing with many different organizations, data products, and with several different types of applications for modeling and analysis. Collaboration between QuakeSim developers and data product providers and downstream consumers early on will result in improved efficiency across the board for earthquake studies and analysis of geodetic imaging observations.

REFERENCES

[1] Donnellan, A., B. H. Hager, and R. W. King, Discrepancy between geologic and geodetic deformation rates in the Ventura basin, Nature, 366, 333–336, 1993.

[2] Pierce, M., Gao, X., Pallickara, S. L.,Guo, Z., and Fox, G., The Quakesim portal and services: new approaches to

science gateway development techniques. Concurrency and Computation: Practice and Experience 22(12): 1732-1749 (2010).

[3] Donnellan, A, J Parker, M Glasscoe, E De Jong, M Pierce, G Fox, D McLeod, J Rundle, L Grant Ludwig, A Distributed Approach to Computational Earthquake Science: Opportunities and Challenges, Computing In Science And Engineering, 2012.

[4] Wang, J., M. Pierce, Y. Ma, G. Fox, A. Donnellan, J. Parker, and M. Glasscoe, Using Service-Based GIS to Support Earthquake Research and Disaster Response, *Computing in Science and Engineering*, 14, 21-30, September 2012.

[5] Kai Hwang, Geoffrey Fox, and Jack Dongarra, Distributed and Cloud Computing: from Parallel Processing to The Internet of Things. 2011: Morgan Kaufmann Publishers.

[6] Jeffrey Dean and Sanjay Ghemawat, MapReduce: simplified data processing on large clusters. Commun. ACM, 2008. 51(1): p. 107-113. DOI: http://doi.acm.org/10.1145/1327452.1327492

[7] Yingyi Bu, Bill Howe, Magdalena Balazinska, and Michael D. Ernst, HaLoop: Efficient Iterative Data Processing on Large Clusters, in The 36th International Conference on Very Large Data Bases. September 13-17, 2010, VLDB Endowment: Vol. 3. Singapore. http://www.ics.uci.edu/~yingyib/papers/HaLoop_camera_ready.pdf.

[8] 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, http://grids.ucs.indiana.edu/ptliupages/publications/hpdc-camera-ready-submission.pdf.

[9] Matei Zaharia, Mosharaf Chowdhury, Michael J. Franklin, Scott Shenker, and Ion Stoica, Spark: Cluster Computing with Working Sets, in 2nd USENIX Workshop on Hot Topics in Cloud Computing (HotCloud '10). June 22, 2010. Boston. http://www.cs.berkeley.edu/~franklin/Papers/hotcloud.pdf

[10] Geoffrey C. Fox, Data intensive applications on clouds, in Proceedings of the second international workshop on Data intensive computing in the clouds. 2011, ACM. Seattle, Washington, USA. pages. 1-2. DOI: 10.1145/2087522.2087524.

[11] Jackson, R. Lavanya Ramakrishnan, Karl J. Runge, and Rollin C. Thomas, "Seeking supernovae in the clouds: a performance study," *Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing*. ACM. Chicago, Illinois. 421-429. DOI: 10.1145/1851476.1851538. (2010).

[12] Mehrotra, P., J. Djomehri, S. Heistand, R. Hood, H. Jin, A. Lazanoff, S. Saini, and R. Biswas. "Performance evaluation of Amazon EC2 for NASA HPC applications". in Proceedings of the 3rd workshop on Scientific Cloud Computing Date 2012: ACM.

[13] Saini, S., S. Heistand, H. Jin, J. Chang, R. Hood, P. Mehrotra, and R. Biswas. "An Application-based Performance Evaluation of NASA's Nebula Cloud Computing Platform". in High Performance Computing and Communication & 2012 IEEE 9th International Conference on Embedded Software and Systems (HPCC-ICESS), 2012 IEEE 14th International Conference on 2012

[14] Thilina Gunarathne, Bingjing Zhang, Tak-Lon Wu, and Judy Qiu, "Scalable Parallel Computing on Clouds Using Twister4Azure Iterative MapReduce" Future Generation Computer Systems 2012. To be published.

[15] Geoffrey Fox and Dennis Gannon "Using Clouds for Technical Computing" Technical Report January 2 2013 http://grids.ucs.indiana.edu/ptliupages/publications/Clouds_Technical_Computing_Review.pdf

[16] Knight, D.; Shams, K.; Chang, G.; Soderstrom, T.; , "Evaluating the efficacy of the cloud for cluster computation," Aerospace Conference, 2012 IEEE , vol., no., pp.1-10, 3-10 March 2012 doi: 10.1109/AERO.2012.6187359

[17] Humphrey, M., N. Beekwilder, J.L. Goodall, and M.B. Ercan, "Calibration of Watershed Models using Cloud Computing."

[18] Chen, Z., N. Chen, C. Yang, and L. Di, Cloud Computing Enabled Web Processing Service for Earth Observation Data Processing. Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of, 2012. 5(6): p. 1637-1649. DOI:10.1109/jstars.2012.2205372

[19] Humphrey, M., Z. Hill, K. Jackson, C. van Ingen, and Y. Ryu. Assessing the value of cloudbursting: A case study of satellite image processing on windows azure. in E-Science (e-Science), 2011 IEEE 7th International Conference on 2011

[20] Li, J., M. Humphrey, D. Agarwal, K. Jackson, C. van Ingen, and Y. Ryu. escience in the cloud: A modis satellite data reprojection and reduction pipeline in the windows azure platform. in Parallel & Distributed Processing (IPDPS), 2010 IEEE International Symposium on 2010

[21] Agarwal, D. and S.K. Prasad. Lessons Learnt from the Development of GIS Application on Azure Cloud

Platform. in Cloud Computing (CLOUD), 2012 IEEE 5th International Conference on 2012: IEEE.

[22] Agarwal, D., S. Puri, X. He, S.K. Prasad, and X. Shi, Crayons-a cloud based parallel framework for GIS overlay operations.

[23] Martinez, N.D., P. Tonin, B. Bauer, R.C. Rael, R. Singh, S. Yoon, I. Yoon, and J.A. Dunne, Sustaining Economic Exploitation of Complex Ecosystems in Computational Models of Coupled Human-Natural Networks. Proceedings of The Association for the Advancement of Artificial Intelligence, 2012.

[24] Geoffrey C. Fox, Gregor von Laszewski, Javier Diaz, Kate Keahey, Jose Fortes, Renato Figueiredo, Shava Smallen, Warren Smith, and Andrew Grimshaw "FutureGrid - a reconfigurable testbed for Cloud, HPC and Grid Computing" Technical Report July 13 2012, Chapter in "Contemporary High Performance Computing: From Petascale toward Exascale", editor Jeff Vetter Chapman & Hall/CRC Press

BIOGRAPHIES

Andrea Donnellan is a principal research scientist at NASA's Jet Propulsion Laboratory, and Adjunct Assistant Professor of Research of Earth Sciences at the University of Southern California. She is Principal Investigator of NASA's QuakeSim project, which was co-winner of NASA's Software of the Year Award in 2012. Donnellan was Deputy Manager of the JPL's Science Division, Pre-Project Scientist of an L-band radar mission, and NASA's Applied Sciences Program Area Co-Lead for Natural Disasters. Donnellan received a B.S. in geology from the Ohio State University in 1986, a master's and Ph.D. in geophysics from Caltech in 1988 and 1991 respectively, and an M.S. in Computer Science from the University of Southern California in 2003. She held a National Research Council postdoctoral fellowship at NASA's Goddard Space Flight Center. Donnellan has conducted field studies globally in tectonically active areas, and on ice sheets and glaciers, and has received numerous awards.

Jay Parker joined the Satellite Geodesy and Geodynamics Systems Group at JPL in 1996, and has been part of the JPL technical staff since 1989. He completed both a master's and PhD in Electrical Engineering from the University of Illinois (Urbana-Champaign), and graduated with a Bachelors of Science from the California Institute of Technology in 1981. His professional interests lie in applications of fast and accurate numerical models to geophysical remote sensing. Past modeling projects include vortex formation in the ionospheric D region, parallel supercomputer modeling of radar scattering and antenna power patterns, and high-fidelity modeling of clear-air infrared spectra for determining climate change and pollution sources. He is currently working on methods to invert SCIGN GPS data to determine earthquake and after-slip fault movements, finite element models of earthquake cycles, and new methods for GPS data processing on supercomputers. Jay has been inducted into Tau Beta Pi, and received a JPL Technology and Applications Programs Group Achievement Award. He is a member of the American Geophysical Union, and the IEEE Antennas and Propagation Society.
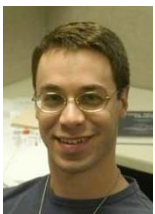
Robert Granat is currently Group Supervisor of the Machine Learning and Instrument Autonomy group at JPL, and has been a member of technical staff there since 1996. He received his M.S. and Ph.D. degrees in Electrical Engineering from the University of California, Los Angeles, and his B.S. from the California Institute of Technology. Since 1999, he has been working on QuakeSim and related projects to perform statistical machine learning based health monitoring, signal classification, and anomaly detection on seismic and GPS sensor networks. His other research interests include autonomous robotic navigation, scalable scientific computing, and radiation fault tolerant algorithms for spaceborne computation.

Maggi Glasscoe is a Geophysicist in the Solid Earth Group at the Jet Propulsion Laboratory, California Institute of Technology. Glasscoe received a B.S. in Geological Sciences and a B.A. in Print Journalism from the University of Southern California in 1997 and a M.S. in Geology from the University of California, Davis in 2003. She has experience working with a number of modeling codes, including viscoelastic finite element models (the JPL developed Geophysical Finite Element Simulation Tool, GeoFEST, in particular). Her interests include disaster response, modeling deformation of the Earth's crust to study postseismic response to large earthquakes, numerical models of the rheological behavior of the lower crust, and simulations of interacting fault systems. She is is a researcher on the QuakeSim project and the Principal Investigator of E-DECIDER (Earthquake Data Enhanced Cyber-Infrastructure for Disaster Evaluation and Response).

Marlon Pierce is the Assistant Director for the Science Gateways Group in Research Technologies Applications at Indiana University. Pierce received his Ph.D. Florida State University (Physics) in 1998 in computational condensed matter physics. His current research and development work focuses on computational sciences with an emphasis on Grid computing and computational Web portals. Prior to forming the Science Gateway Group, Pierce served as assistant director for the Community Grids Laboratory at Indiana University's Pervasive Technologies Institute. Pierce supervises the research activities of software engineering staff and Ph.D. students, and serves as principal investigator on multiple federally-funded research projects. Pierce leads research efforts in the following areas: the application of service-oriented architectures and real-time streaming techniques to geographical information systems and sensor networks; the development of open source science Web portal software for accessing Grid computing and data resources; and Grid-based distributed computing applications in computational chemistry and material science, chemical informatics, and geophysics.

Jun Wang is a GIS specialist in the Community Grids Lab, Indiana University. His current research interests are in the areas of large scale spatial data processing and visualization with Cloud computing technology. He joined QuakeSim team in July 2010, and develops the software components for QuakeSim visualization products.

Geoffrey Fox received a Ph.D. in Theoretical Physics from Cambridge University and is now professor of Informatics and Computing, and Physics at Indiana University where he is director of the Digital Science Center and Associate Dean for Research and Graduate Studies at the School of Informatics and Computing. He previously held positions at Caltech, Syracuse University and Florida State University. He has supervised the PhD of 62 students and published over 600 papers in physics and computer science. He currently works in applying computer science to Bioinformatics, Defense, Earthquake and Ice-sheet Science, Particle Physics and Chemical Informatics. He is principal investigator of FutureGrid – a facility to enable development of new approaches to computing. He is involved in several projects to enhance the capabilities of Minority Serving Institutions.

Dennis McLeod is a Professor of Computer Science at the University of Southern California, and Director of the Semantic Information Research Laboratory. He received his Ph.D., M.S., and B.S. degrees in Computer Science and Electrical Engineering from MIT. Dr. McLeod has published widely in the areas of data and knowledge base systems, federated databases, database models and design, ontologies, knowledge discovery, scientific data management, information trust and privacy, and multimedia information management. His current research focuses on: structured domain ontologies; semantic web; database semantic heterogeneity resolution and inter-database correlation; personalized information management and customization; information management environments for Earth, marine, and climate science; the architecture of data centers providing massive storage via virtualization and data clouds; social networking information management and information trust; and service-based information access and delivery frameworks.

Civil Engineering and Geology at University of California, Davis. His research is focused on understanding the dynamics of earthquakes through numerical simulations; pattern analysis of complex systems; dynamics of driven nonlinear Earth systems; and adaptation in general complex systems. Computational science and engineering is an emerging method of discovery in science and engineering that is distinct from, and complementary to, the two more traditional methods of experiment/observation and theory. The emphasis in this method is upon using the computer as a numerical laboratory to perform computational simulations to gain insight into the behavior of complex dynamical systems, to visualize complex and voluminous data sets, to perform data mining to discover hidden information within large data sets, and to assimilate data into computational simulations. Professor Rundle is a Fellow of the American Geophysical Union.

Eric M. Heien is lead programmer in the Computational Infrastructure for Geodynamics (CIG) group at the University of California, Davis. His research interests include parallel and distributed computing, volunteer computing, and high-performance computing for scientific simulation. He received a PhD in computer science from Osaka University.

Lisa Grant Ludwig is an Associate Professor in Public Health at the University of California, Irvine and former Associate Director of the California Institute for Hazards Research. She earned a Ph.D. in Geology with Geophysics and dual M.S. degrees in Geology and Environmental Engineering from Caltech, and a B.S. from Stanford in Applied Environmental Earth Sciences. Her research interests include natural hazards, paleoseismology, active faults, seismic hazard, environmental health and geology. Her research group addresses natural hazards and disasters from a geologic perspective, with emphasis on earthquakes. The group focuses on defining the potential for large earthquakes, and working collaboratively on developing forecasts, hazard models and effective responses. Results of the work are applied for disaster preparedness planning, structural design, land-use planning, seismic risk assessment and public education about earthquake hazard.

John Rundle is an Interdisciplinary Professor of Physics,