QuakeSim: Enabling Model Interactions in Solid Earth Science Sensor Webs

Andrea Donnellan, Jay Parker, Charles Norton, Gregory Lyzenga, Margaret Glasscoe
Jet Propulsion Laboratory, California Institute of Technology
4800 Oak Grove Drive
Pasadena, CA 91109
818-354-4737
Andrea.Donnellan@jpl.nasa.gov

Geoffrey Fox and Marlon Pierce
Community Grids Laboratory, Indiana University
501 N. Morton, Suite 224
Bloomington, IN 47404

John Rundle
Department of Physics, University of California, Davis
One Shields Avenue
Davis, CA, 95616

Dennis McLeod
University of Southern California
Mail Code 0781, 3651 Trousdale Parkway
Los Angeles, CA 90089

Lisa Grant
University of California, Irvine
Irvine, CA 92697

Walter Brooks
NASA Ames Research Center
Moffett Field, California 94035

Terry Tullis
Department of Geological Sciences, Brown University
Providence, RI 02912

Abstract—QuakeSim is problem-solving environment for understanding earthquake processes through the integration of multiscale models and data. The goal of QuakeSim is to substantially improve earthquake forecasts, which will ultimately lead to mitigation of damage from this natural hazard. Improved earthquake forecasting is dependent on measurement of surface deformation as well as analysis of geological and seismological data. Space-borne technologies, in the form of continuous GPS networks and InSAR satellites, are the key contributors to measuring surface deformation. We are expanding our QuakeSim Web Services environment to integrate both real-time and archival sensor data with high-performance computing applications for data mining and assimilation. We are federating sensor data sources, with a focus on InSAR and GPS data, for an improved modeling environment for forecasting earthquakes. These disparate measurements form a complex sensor web in which data must be integrated into comprehensive multi-scale models. In order to account for the complexity of modeled fault systems, investigations must be carried out on high-performance computers. We are building upon our "Grid of Grids" approach, which included the development of extensive Geographical Information System-based "Data Grid" services. We are extending our earlier approach to integrate the Data Grid components with improved "Execution Grid" services that are suitable for interacting with high-end computing resources. These services are being deployed on the Columbia computer at NASA Ames and the Cosmos computer cluster at JPL. [1] [2]

TABLE OF CONTENTS

1———————————————

[1] 1-4244-0525-4/07/$20.00 ©2007 IEEE.
[2] IEEEAC paper #1039, Version 1, Updated October 23, 2006

1. INTRODUCTION

For over a decade, the solid Earth science community has been analyzing data from radar satellites and preparing for a US-led Interferometric Synthetic Aperture Radar (InSAR) mission. Other spaceborne data used in modeling earthquake processes come from GPS from the Southern California Integrated GPS Network (SCIGN), a network of 250 continuously operating GPS stations in southern California, and the Plate Boundary Observatory, covering a broader region of the western US. GPS data from SCIGN and PBO are in the form of daily position time series of surface deformation. The data can be analyzed for transient station motions from the time series, or the long-term vector motions can be used to model fault activity in southern California.

A goal of this project is to prepare for the anticipated data deluge from an InSAR mission as well as to enable fusion and analysis of data from different sensors and sources. This will serve to help develop the necessary data and modeling infrastructure for developing an improved understanding of earthquake processes.

Use of NASA's Columbia computer is enabling us to construct the complex and detailed models necessary for accurate understanding and forecasting of earthquakes. Columbia provides an integrated computing, visualization and data storage environment. This project will integrate spaceborne data with ground-based data and simulations using high-end computational infrastructure. The focus is to meet a practical need of contributing to the mitigation of disasters from earthquakes while developing Web Service and computational tools that can be applied more broadly to other scientific problems. We are addressing issues such as security, communication with supercomputers through a Web Services environment, distributed data sources, and data from upcoming missions.

2. RELEVANCE

Enabling model interactions in sensor webs should lead to improved estimates of damage probability for mitigating losses from potential earthquakes. Our sensor web data sources consist of surface deformation data from GPS and InSAR satellites and seismicity from the California Integrated Seismic Network. These are coupled to QuakeTables, the earthquake fault database service and simulation and data-mining tools (Figure 1) through execution services for application management. The high performance software GeoFEST, Virtual California, and PARK are used for simulations and the PI (Pattern Informatics) and RDAHMM (Regularized Deterministic Annealing Hidden Markov Models) methods make up the data-mining tools. Outputs of these models describe potential earthquakes. The locations of potential earthquakes, combined with attenuation models, indicating

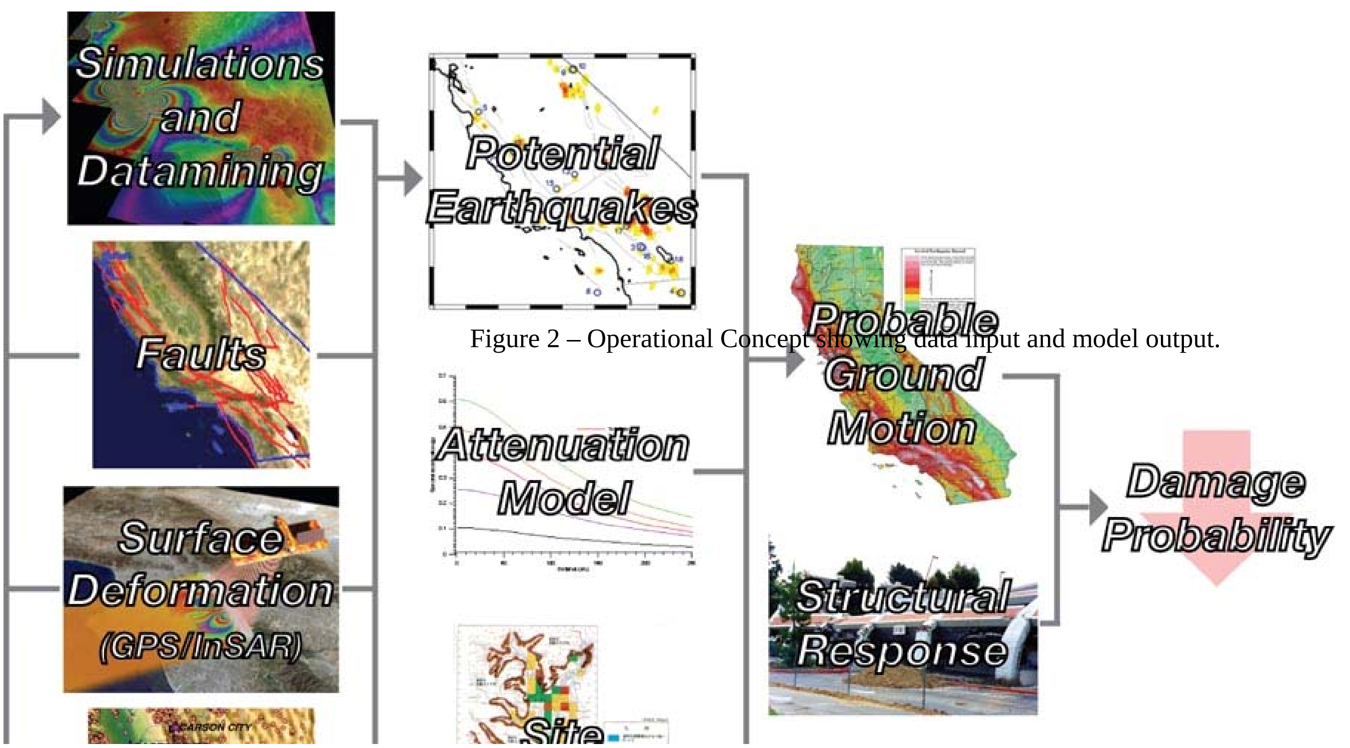
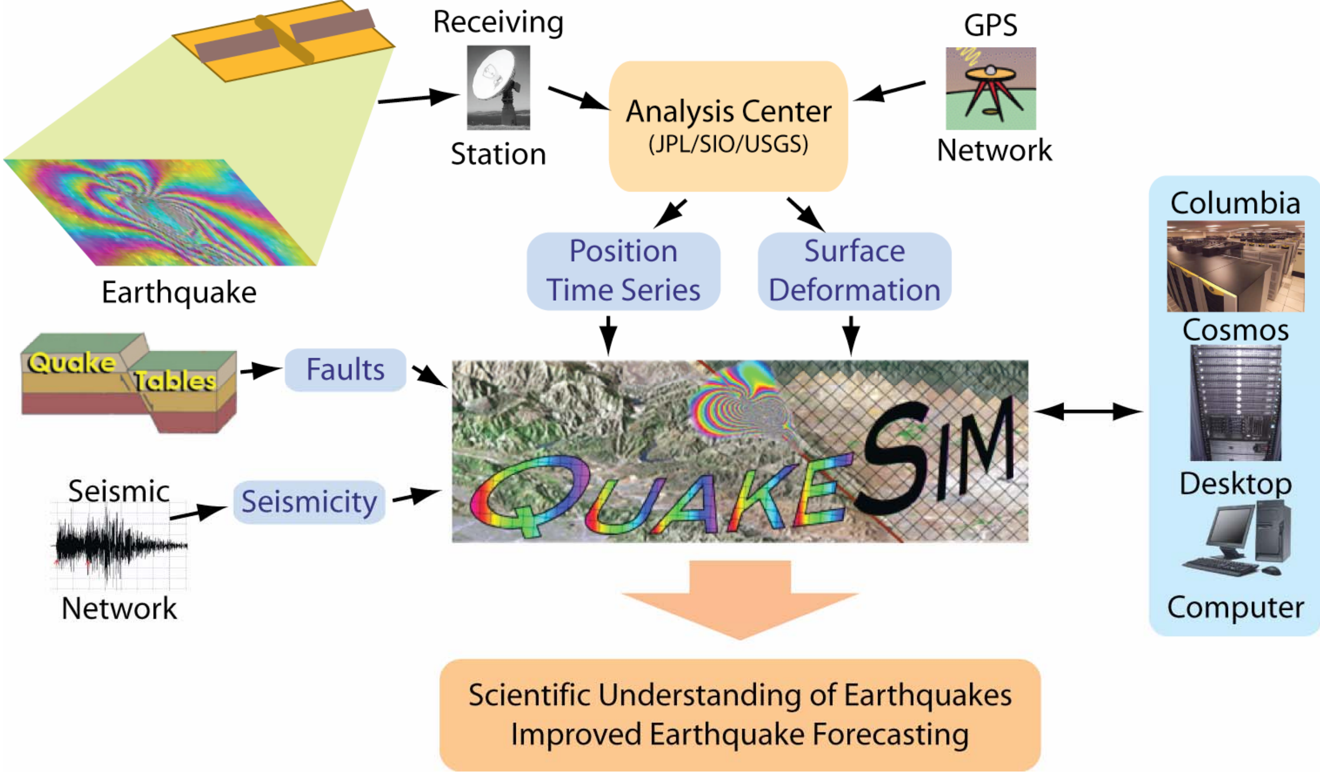


Figure 2 – Operational Concept showing data input and model output.

Figure 1 – Risk estimation processing showing the relation between a geophysical sensor web and earthquake risk mitigation. The sensor web is on the left column and feeds into simulation and data mining tools. These in turn flow down and combine with other inputs to produce an estimate of damage probability.



how shaking diminishes from the source of the earthquake, and local site effects are used to determine probable ground motions. These potential ground motions combined with the response of structures are used to estimate damage probability. Once this is known, strategies to mitigate the effects of earthquakes such as targeted retrofitting and disaster preparation can be more effectively developed.

3. Technology

QuakeSim focuses on ingesting data and models into simulations run on desktop workstations or on high-end computers. Such a system requires interoperable data ingestion. We are using emerging grid and web common languages and semantic metadata to enable data exchange and fusion. QuakeSim involves data assimilation, high performance computing, semantic data modeling, and constructing Geographical Information Systems as Web Service-based Grids. The current objectives of QuakeSim are a) the extension of data assimilation and analysis applications to high performance computing platforms, b) the support for InSAR data in anticipation of future InSAR satellite missions, and c) and the deployment and integration of services for high performance computing resources with our data services infrastructure. QuakeSim provides interoperable ingestion of data as well as easy plug-and-play structure for scientific algorithms, using emerging grid and web common languages. The goal is to rapidly produce scientific understanding of earthquake processes and develop testable earthquake forecasts and forecasting methodology to be integrated into operational agencies (Figure 2).

QuakeSim focuses primarily on the interseismic process. In other words, this project does not involve the shaking part, or waveforms associated with earthquakes. We are studying and modeling the strain and fault interactions that are associated with or otherwise produce large damaging earthquakes. Surface deformation data, largely in the form of spaceborne data from GPS and InSAR, are key inputs to studying these processes. We are, therefore, focusing on ingesting InSAR data and integrating it with other data types such as GPS and paleoseismic (fault) data into models of interacting fault systems. These data sources are made available through both archival and real-time Web Services based on the Geographical Information System service specifications defined by the Open Geospatial Consortium. A Web Services management system, HPSearch, and Web Service information systems (based on UDDI and WS-Context) provide the glue that couples these data sources and the applications.

The QuakeSim web services environment currently does not support high performance computing resources, which must be used to realistically investigate the more interesting data mining and assimilation problems. Thus the two primary challenges at present are a) building upon our earlier successes in high performance computing applications by applying them to new and larger data sets, and b) integrating the high performance applications and their associated sensor data (Table 1) with our Service Oriented Architecture.

Under previous work [1] we have developed component-based portals and Web Services to support geophysical applications. Applications include traditional high performance software as well as data analysis and assimilation codes. The modeling applications include GeoFEST [2], a finite element model that simulates stresses associated with earthquake faults, Virtual California [3], which simulates large, interacting fault systems, and PARK [4], which simulates complete earthquake cycles and earthquake interaction. Analysis methods include Pattern Informatics [5], which examines seismic archives to forecast geographic regions of future high probability for intense earthquakes, and RDAHMM [6], a time series analysis application that can be used to determine state changes in instrument signals (such as generated by Global Positioning System arrays).

Table 1 – Sensor data and associated software

Sensor Data Type	Software	Description
InSAR	GeoFEST	Models surface deformations caused by faults stress; directly comparable to InSAR results.
Seismic activity records (SCSN, SCEDC, etc).	Virtual California	Uses interacting fault models, calculates long range earthquake activity forecasts and compares to seismic activity archives for best-cost analysis.
Seismic activity records	Pattern Informatics	Hot-spot forecasting based on data assimilation of seismic activity archives.
GPS position archives (JPL, SOPAC, etc)	RDAHMM	Time series analysis and mode detection in GPS and other signals.
GPS, InSAR	Simplex	Optimally finds a dislocation model of fault slip that accounts for GPS and inSAR deformation data.
Seismicity patterns	PARK	Determines model parameters that best reproduce the observed seismicity patterns.

GIS Services, Sensor Webs, and Workflow

Much of our Web Service development work has focused on developing Geographical Information System (GIS) data services and integrating them with the applications listed above. We have implemented and extended a number of the Open Geospatial Consortium's standards, including the Web Feature Service [7], a general purpose data archive, typically built on the top of a relational database, and the Web Map Service, a Web Service for rendering Web Feature Service entries as human readable maps. The work summarized at *www.crisisgrid.org.* An extensive technical report is available [8].

Our Web Feature Service is capable of storing archived data (such as GPS, seismic records, and fault data), but we are also integrating real-time GPS data streams with applications. Our approach is based on combining the Sensor Web Service modeling languages such as SensorML [9] with topic-based publish/subscribe middleware that can be used to route and filter live data streams. We have prototyped this using RDAHMM (Regularized Deterministic Annealing Hidden Markov Model), a statistical time series analysis code applied to seismicity and position time series data. We currently support over 7 GPS sub-networks that contain over 70 individual stations in the Southern California Integrated GPS Network. We are interested in addressing scaling and performance issues as we increase the number of GPS stations and in combining the streaming applications with other event detection applications.

Both archival and real-time Grid applications require the integration of many different services (for data, execution, stream routing, etc.) into a complex meta-application. The management of these services is typically called workflow. We have developed a general-purpose Web Services management tool, HPSearch (www.hpsearch.org), which can manage SERVO's Geographical Information System services, GPS data streams, and applications such as Pattern Informatics and GeoFEST [10]. In this project, we propose to extend HPSearch to support secure Web Services (such as is available from Globus) for running applications on high performance computers.

We have also developed distributed information services to support both stateless and stateful information in distributed Web Services. These are based on the UDDI and WS-Context specifications, respectively. Stateless information sources are those that respond to all queries in the same way. Stateful information depends on which client makes the query. For example, different users may be interested in receiving information about different collections of GPS stations. Matching these users to station streams is an example of a context [11].

The challenge of this project is to combine sensor data, accessible through the Data Grid Services with high performance resources to investigate challenging problems in earthquake modeling and forecasting using the applications listed in Table 1. Our philosophy is to not reinvent technology but will instead rely upon third party solutions. The current version of the Globus Toolkit (Version 4) [12] has been significantly revised to make use of Web Service standards. Also, the client-programming environment for Grid portals (grid portlets [13; 14] and the Java CoG Kit [15]) has undergone significant enhancements. Globus services solve the two shortcomings in our current job management services: they have built-in job managers that translate user requests into different job scheduling systems (including PBS and others), and they have a security system that is capable of spawning jobs as specific users.

4. APPLICATIONS

Virtual California and GeoFEST

Virtual California (VC) is a numerical simulation code for the system-level dynamics of the vertical strike-slip fault configuration in California [16;17]. The majority of plate boundary deformation in California is accommodated by slip (i.e. earthquakes) on the strike-slip faults included in VC models. We have recently developed a much more detailed model having $12{,}288 = 3*2^{12}$ elements at a scale of resolution of approximately 3 km. We have recoded in C++ and parallelized the code with an efficient MPI implementation using MPI-II. A limitation of Virtual California is that it currently models only vertical strike slip faults, but this will change soon. We are adding dipping thrust and normal faults to the existing QuakeTables fault data base, and will couple Virtual California with GeoFEST to compute the necessary Green's functions. In the latest version, VC forecasting calculations can be trained on existing seismic data archives. The 1994 Northridge earthquake on a blind thrust fault beneath the Los Angeles metropolitan region demonstrated the importance of adding these faults. Dipping faults must be included in the database to enable forecasting based on VC simulations and interpretation of hotspot maps based on processing sensor web archives of seismicity data.

PARK

The PARK code [4] runs on both Columbia and the JPL cluster. It uses fast multipoles to allow the efficient use of many elements in a boundary element representation of earthquake slip using rate and state friction and radiation damping, the most accurate representation presently available of the quasi-dynamic behavior of faults. Ongoing studies using the PARK code are focused on the Richter magnitude ranges from 1.0 to 6.0 at Parkfield California, using a multi-scale grid representation. The same approach using the multipole method and a wide range of element sizes can be used on any array of fault elements on any number of faults, covering any sized area of interest – our Virtual California code is an example of a coarser scale gridding scheme covering a much larger area and using a simplified frictional representation.

PI and RI Methods

We have developed two types of hotspot map, the original seismic Pattern Informatics (PI) map and the Relative Intensity (RI) map. Both were first discussed as possible forecast tools over a 10-year time span in [18]. In that work the original hotspot map was published February 19, 2002. During the time of the forecast (January 1, 2002 – 2010), 19 significant events with M ~ 5.0 or greater have occurred in the area of the original forecast map, with 17 of them showing a close spatial relationship to the colored hotspots. 16 of these events occurred after February 19, 2002, and 14 of those were among the group near the hotspots. More recently, we have shown [19] that it is possible to develop an ensemble classifier schema using both RI and PI ("RIPI") that allows us to determine whether a broad geographic region ("northern" or "southern" California) is currently in a "high risk" or "low risk" state for major earthquakes having M > 6.0). We find that currently, northern California is in a high risk state, whereas southern California is in a low risk state.

RDAHMM

Signals of interest, particularly those indicating stress transfer between faults, are very subtle, and are often overlain by other sorts of signals, arising from sources as diverse as aquifer activity and atmospheric disturbances. The statistical modeling approach, RDHAMM [6], allows us to automatically infer modes of activity within individual time series and across a network of sensors. Currently the method is applied to classification of seismicity data and position time series data from GPS. The unique modeling technology allows us to be effective even in cases in which there is no model for the observed system, as well as overcome stability problems that plague standard methods.

One challenge we face is that the method needs to be computationally swift enough to be applied in real time to streaming sensor data. Current model fitting methods, including the one we outline above, are iterative approaches that can take an unacceptably long time to converge. We are solving this problem by using methods such as conjugate gradient acceleration to speed convergence. In order to take advantage of situations in which multiple computational processors are available, wherever possible we are implementing our methods so that they can be run in parallel.

Data Fitting Techniques

In order to develop an effective and precise algorithmic component for the coming complex sensor web, we need to integrate data and modeling software at the level where sensor observations are sensitive to model parameters. Because we intend the system to be used where processes are still uncharted, it is essential that new modeling components be straightforward to add, by geophysicists beyond the current team. We will plan and build a data-fitting core software module such that new data or new models may be correctly combined with prior data using all the information in each, fully compatible with distributed components.

Initial progress toward this will be rapid by expanding the **Simplex** application of QuakeSim, to include additional physical deformation processes and data types, and to separate the core data-fitting functions from the modeling software and the data input. Included processes will include subsurface loads (aquifers, oil extraction, and volcanic inflation), additional fault models such as locking depth slip, addition of dipping faults, and partially known systematic errors. The QuakeSim environment will be enhanced for ease of problem setup and documentation for these process estimation runs. Simplex estimates slip on faults and will estimate other deformation processes; but it is limited to homogeneous elastic models, and so we will develop a correction step, wherein the Simplex estimate of fault slip automatically becomes the basis of a suite of detailed finite element simulation including known materials variation. This suite of results will indicate the best first-order correction to the fault parameters, often a sufficient result. The potential for fully-iterative optimal fits to data will be evaluated.

The goal is a broadly adaptable integrated system of precision surface deformation monitoring, combined with a modeling system that incorporates processes at multiple scales. This will allow definition of a baseline model of regional and global deformation processes, which can be continuously compared with sensor observations for automatic early detection of unusual events.

To account for the relevant data (e.g. InSAR, GPS) necessary for the models and to constrain the simulations we are developing a data-fitting core software module that connects in a standard way with separate modeling and observational models that incorporate realistic instrumental errors and covariance, capable of running on remote machines. The core module would be dedicated to least-squares fitting of all data and models chosen by the user, minimizing a chi-square cost function that measures the weighted deviation of the observed data to the sum of the effects of all participating modeling modules. Other kinds of computations, including registering observations in a global reference frame, atmospheric and orbital effects, and geophysical processes will be relegated to the modeling modules. The end result makes optimal use of all kinds of solid earth deformation sensors and will be extensible to observations such as gravity. It provides a framework in which various models may be steadily and systematically improved by an ongoing community effort, and various sensors can be incorporated directly. It will also lead to a slow-deformation background model that will form the baseline from which unexpected deviations can be rapidly detected and classified.

5. CONCLUSIONS

Current earthquake risk estimation is based on static models inferred from past earthquake activity as determined through paleoseismology and historical earthquakes. Measurement of crustal deformation and rates of strain contribute to earthquake hazard assessment, largely because of measurement of surface deformation, largely through GPS and InSAR. Earthquake fault systems are continuously changing state based on deformation of the Earth’s crust and mantle as well as strain release and transfer from earthquakes. It is important to develop time-dependent models for earthquake forecasting.

QuakeSim aims to integrate both real-time and archival sensor data with high performance applications for data mining and assimilation. In order to achieve the development of complex models of interacting fault systems, models are coupled together and users access those models through a web services environment. QuakeSim integrates distributed heterogeneous data sources through federated databases, and our goal is to carry out models and simulations on high performance or appropriate computers through our grid of grids approach.

References

[1] Donnellan, A., J. Rundle, G. Fox, D. McLeod, L. Grant, T. Tullis, M. Pierce, J. Parker, G. Lyzenga, R. Granat, M. Glasscoe, QuakeSim and the Solid Earth Research Virtual Observatory, *PAGEOPH*, in press.

[2] Parker, Jay, Andrea Donnellan, Gregory Lyzenga, John B. Rundle, Terry Tullis: Performance Modeling Codes for the QuakeSim Problem Solving Environment. International Conference on Computational Science: 855-862, 2003.

[3] Rundle, J.B., K.F. Tiampo, W. Klein and J.S.S. Martins, Self-organization in leaky threshold systems: The influence of near mean field dynamics and its implications for earthquakes, neurobiology and forecasting, *Proc. Nat. Acad. Sci*. USA, **99**, Supplement 1, 2514-2521, (2002).

[4] Tullis, T. E., Earthquake models using rate and state friction and fast multipoles, Geophysical Research Abstracts (European Geophysical Society), 5, 13211, 2003.

[5] Tiampo, K.F., Rundle, J.B., McGinnis, S., and Klein, W. Pattern dynamics and forecast methods in seismically active regions, Pure and Applied Geophysics, 159, 2002.

[6] Granat, R. A., *Regularized Deterministic Annealing EM for Hidden Markov Models*, Ph.D. Thesis, University of California, Los Angeles, 2004.

[7] Vretanos, P (ed.), *Web Feature Service Implementation Specification*, OpenGIS project document: OGC 02-058, version 1.0.0, 2002.Zebker, H., J. Andrews, Y. Fialko, M. Glasscoe, B. Holt, K. Jezek, I. Joughin, Z. Lehner, B. Minster, P. Rosen, J. Rundle, J. Sauber, M. Simons, W. Thatcher, InSAR Workshop Summary Report, Oxnard, California, 64 pp., October 20–22, 2004.

[8] Aktas, Mehmet, Galip Aydin, Geoffrey Fox, Harshawardhan Gadgil, Marlon Pierce and Ahmet Sayar, “SERVOGrid Technical Documentation” NASA AIST Technical Report.

[9] Botts, M (ed), OpenGIS Sensor Model Language (SensorML) Implementation Specification, OGC 05-86r2, 2006.

[10] Aydin, Galip, Mehmet S. Aktas, Geoffrey C. Fox, Harshawardhan Gadgil, Marlon Pierce, Ahmet Sayar “SERVOGrid Complexity Computational Environments (CCE) Integrated Performance Analysis.” In proceedings of GRID2005, Seattle Washington., November 12-13, 2005.

[11] Aktas, Mehmet, Geoffrey C. Fox, Marlon Pierce “An Architecture for Supporting Information in Dynamically Assembled Semantic Grids” Technical report August 2005, accepted for publications in Proceedings of Semantics, Knowledge, and Grid, Beijing, China, November 2005.

[12] Foster, I, C. Kesselman, G. Tsudik, S. Tuecke, A Security Architecture for Computational Grids. Proc. 5th ACM Conference on Computer and Communications Security Conference, pp. 83-92, 1998.

[13] Alameda, Jay, Marcus Christie, Geoffrey Fox, Joe Futrelle, Dennis Gannon, Mihael Hategan, Gopi Kandaswamy, Gregor von Laszewski, Mehmet A. Nacar, Marlon Pierce, Eric Roberts, Charles Severance, and Mary Thomas, "The Open Grid Computing Environments Collaboration: Portlets and Services for Science Gateways." Submitted to "Science Gateways" special issue of Concurrency and Computation: Practice and Experience, 2006.

[14] Nacar, Mehmet, Mehmet Aktas, Marlon Pierce, Zhenyu Lu, Gordon Erlebacher, Dan Kigelman, Evan F. Bollig, Cesar De Silva, Benny Sowell, and David A. Yuen. "VLab: Collaborative Grid Services and Portals to Support Computational Material Science" GCE'05 Workshop on Grid Computing.at SC05 Seattle, WA. November 18 2005.

[15] Amin, Kaizar, Mihael Hategan, Gregor von Laszewski, Nestor J. Zaluzec: Abstracting the Grid. PDP 2004: 250-257.

[16] Rundle, J.B., D.L. Turcotte, and W. Klein eds., *Geocomplexity and the Physics of Earthquakes*, Geophysical Monograph 120, American Geophysical Union, Washington, DC, 2000.

[17] Rundle, P.B., J.B. Rundle, K.F. Tiampo, J.S.S. Martins, S. McGinnis and W. Klein, Nonlinear network dynamics on earthquake fault systems, *Phys. Rev. Lett*., **87**, 148501, 2001.

[18] Rundle, J.B., K.F. Tiampo, W. Klein and J.S.S. Martins, Self-organization in leaky threshold systems: The influence of near mean field dynamics and its implications for earthquakes, neurobiology and forecasting, *Proc. Nat. Acad. Sci*. USA, **99**, Supplement 1, 2514-2521, 2002.

[19] Holliday, JR, JB Rundle, DL Turcotte, W Klein, KF Tiampo, and A. Donnellan, Space-time correlation and clustering of major earthquakes, *Phys. Rev. Lett.*, in press, December 1, 2006.

BIOGRAPHY

Andrea Donnellan is the Interferometric Synthetic Aperture Radar (InSAR) Science Lead at NASA's Jet Propulsion Laboratory and is a research professor at the University of Southern California. Donnellan uses GPS and InSAR satellite technology coupled with high performance computer models to study earthquakes, plate tectonics, and the corresponding movements of the earth's crust. She has been a geophysicist at JPL since 1993. She received a bachelor's degree from the Ohio State University in 1986, with a geology major and mathematics minor. She received her master's and Ph.D. in geophysics from Caltech's Seismological Laboratory in 1988 and 1991 respectively. Donnellan received an M.S. in Computer Science from the University of Southern California in 2003.