QuakeSim lessons for NASA Earth Science sensor webs

Jay Parker, Andrea Donnellan
Jet Propulsion Laboratory

Marlon Pierce
Crisis Grids Laboratory, Indiana University

Within 5-10 years, there will be a radical shift in the way solid earth science relies on diverse distributed sensors, applications that perform pattern analysis and physical modeling, and investigators collaborating across the internet. While much fundamental research will still take place within traditional methods that rely on a single data archive and local computing facilities, the urgency of understanding natural hazards and their context of the global tectonic system will impel increasing amounts of resource and highly skilled people into new ways that use remote data and computing to assess situations with unprecedented speed and depth of insight. Therefore one of the most urgent needs is for developing technology for using internet-aware tools to find and deliver remote data to high-performance modeling environments.

Background:
Our AIST Sensor Webs project "QuakeSim: Enabling Model Interactions in Solid Earth Science Sensor Webs" addresses the program Topic 3, *Enabling Model Interactions in Sensor Webs*. Our approach is to understand and adapt diverse solid-earth resources now connected to the Internet (data archives, real-time data sources, applications on workstations, applications on supercomputers) to a service-oriented environment, and build up tools in that environment that multiply the benefits of accessibility and collaboration in the context of genuine scientific research.

The new abilities under development will likely form part of the lasting infrastructure of the emerging System of Systems architecture, including sensor webs with the full generality of wireless communications and autonomous analysis and deployment. But we are developing useful components within the current world of data archive centers and monolithic applications.

Data sources:
We will federate sensor data sources for an improved modeling environment for forecasting earthquakes. These will include deformation sensors (GPS, archived InSAR), seismic events, and fault data. Improved earthquake forecasting is dependent on measurement of surface deformation as well as analysis of geological and seismological data.

Models and interactions:
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. The past work of the QuakeSim project has resulted in several promising

modeling and pattern information tools that form the basis of our data processing and science extraction, which along with ingestion of remote data form the proving ground for our sensor web capabilities. These are summarized in Table 1, organized by the most relevant sensor data type:

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.

Table 1. Description of high-performance and other software modules and their relevant sensor data types.

The modeling applications include GeoFEST [Parker, 2003], a finite element model that simulates stresses associated with earthquake faults, Virtual California [Rundle, 2002], which simulates large, interacting fault systems, and PARK [Tullis, 2003], which simulates complete earthquake cycles and earthquake interaction. Analysis methods include Pattern Informatics [Tiampo, 2002], which examines seismic archives to forecast geographic regions of future high probability for intense earthquakes, and RDAHMM [Granat, 2004], a time series analysis application that can be used to determine state changes in instrument signals (such as generated by Global Positioning System arrays).

Grid services:

We are building upon our "Grid of Grids" approach investigated through previous AIST funding. Data and computational resources are packaged or tied to filter layers that provide appropriate metadata and controls for access by web-service tools. Current work includes middleware support for streaming GPS data and topic-based publication and subscription to filters. Output filters include the Open Geospatial Consortium's "Observations and Measurements" schema (GML-OM). Similar wrapping makes the RDAHMM (pattern analysis tool) available for real-time data analysis. Once resources are prepared as Grid tools, workflow becomes manageable, either by special tools (like Crisis Grid Lab's HP-Search) or by simple scripting. We are also building Geographical Information System-based "Data Grid" services to support ontologically described geographic data archives as well as middleware to support real-time streaming GPS data.

In this project we will extend our earlier approach to integrate the Data Grid components with improved "Execution Grid" services that are suitable for interacting with high-end computing resources. By using the latest version of the Globus Toolkit we have a straightforward path to managing job control services and secure access. These services will be deployed on the Columbia computer at NASA Ames and the Cosmos computer cluster at JPL. We plan to explore several use-case models that will require such access. It should be noted that if an entire use case runs on a single computer, Grid access is probably unwarranted; a single large simulation (one common use of high-end computers) doesn't require automated job submission. On the other hand, comparison of recent data with a catalog of simulation results would be a likely early sensor-web use of high-end resources: when the simulations are run, there would be need for efficient job submission control (which might need to be done by remote system oversight). If the bulky results are archived within the secure computing facility, an autonomous alert system detecting a current dangerous situation could need to send an updated observation picture and spawn light-weight comparison jobs to find the best match scenario. Ultimately local sets of sensors may need to have two-way communication with a high-performance data assimilation simulation. For example there would be savings by having remote data filtered by access to the sensitivities and common modes of a high-performance simulation, followed by submission of the cleaned data to the steered simulation process.

Incorporating data and models:

There are many ways to combine data and models, and we are using several to establish methods of general benefit. Perhaps the most primitive is to compare one or a suite of physical model results to data; the web-based GIS tools bring considerable flexibility and power to this task. A second is to take collected and streaming data, for example in the form of multiple-component, multiple source time series, and use pattern information to find special times or hidden states. Some applications (such as Pattern Informatics and RDAHMM) then project those states forward into a forecast. A very simple dislocation model is included in Simplex, and can directly invert observed spatial deformation data to determine hidden fault slip. The model of locally interacting unstable fault patches in PARK produces patterns of seismicity that compare with observed seismicity. Finally Virtual California is a system-wide evolving model. Estimation methods such as least-squares scoring is used to revise the model to fit ongoing seismic and deformation data. Note the best data-fitting methods require combination of data with reported noise level. In some cases this can be inferred by collection system characteristics, but generally the noise must be reported with the data and accessed together.

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. The Simplex code, by virtue of its simplicity, is an ongoing testbed for building plug-and-play data and model components. The code interface has been redesigned, so that data and model components can be read from XML descriptions and combined from separate input files. Data and physical model components link with a data-fitting 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.

Soon our other modeling software will be extended to do comparisons, inversions, and assimilations with a wide range of data types and modeling features. This will enable 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. Such modeling and anomaly detection will drive sensor deployment and data collection at times of high danger or interest.

Contribution to sensor web ideas:
Some very general sensor web concepts suggested by our work so far include:

1) Use web-service oriented metadata and filters for real-time and archived data sources, so that everything becomes available to Internet-aware tools and Grid services.
2) Build generally useful Grid services, incorporating stable third-party components as available. Using such components avoids reinventing the wheel and leverage enormous development work going on in the commercial Web.
3) Use a topic-based publish/subscribe approach for handling real-time data streams, as we have begun to do with streaming GPS.
4) Deploy AI-type pattern analysis software early, as the beginning of autonomous analysis and sensor-web steering.
5) Include high-end computing resources as Grid service applications, while accommodating the need for secure access. Begin discussions of this with computing centers early.
6) Develop high-fidelity physical models that can match observational data, including ability to read metadata and data in formats suitable for internet-based access (such as XML).
7) Support by design, and encourage by all means the inclusion of uncertainties and noise models along with all measurable quantities, both at data sources and in models.
8) Move toward making modeling components into modules that can be combined and reused in several ways. Aquifer depletion and fault motion may have effects that add linearly to deformation; local models may need to be coupled with continental-scale processes.
9) Develop all these parts with active participation of active domain investigators.

Acknowledgements:
Includes research which was performed at the Jet Propulsion Laboratory, California Institute of Technology, under contract to the National Aeronautics and Space Administration.

References:
Granat, R., and A. Donnellan, A hidden Markov model tool for geophysical data exploration, *Pure and Appl. Geophys*, 2271–2284, 2002.

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.

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).

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.

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