Introduction: Driven threshold systems are some of the most important, but least-understood systems in nature. They include earthquakes [1], the human brain [2], convective circulations in the atmosphere [3], one-dimensional electron waves in solids [4], driven foams [5], and magnetic de-pinning transitions in high temperature superconductors [6]. We propose to develop a new information-technology and statistical-physics based approach to understanding the *often unobservable* dynamics in this class of research domains. We will focus on understanding the *observable* space-time patterns that these systems display, using new (GeoInformatics) data-mining, pattern recognition, and ensemble forecasting techniques appropriate for these systems with multiple time and spatial scales. Our specific focus will be earthquake fault systems, which have recently emerged as a paradigm for all driven threshold systems. The special features of data mining, artificial intelligence and other computational techniques for the study of complex systems [7] require that we develop new approaches to Problem Solving Environments (PSE). These must support distributed real time data streams integrated with simulations and pattern analyses. Our PSE work is based on dynamic self defining objects to encapsulate data and simulations and will focus on the earthquake research domain but the ideas, methods, and software produced from study of the various crossover field technologies will support any research domain, which can be abstracted as a similar driven multiscale threshold system with real-time data. We will also pursue a significant component of outreach of our results to research colleagues and the general public using staff at the University of California at Davis and at the University of Colorado.

***Rationale*:** Earthquakes have great scientific, societal, and economic significance. During the past three months alone, the January 13, 2001 magnitude 7.6 El Salvador earthquake, the January 26, magnitude 7.9 Gujarat India earthquake, and the February 28, 2001 magnitude 6.8 Seattle, Washington USA event have killed thousands of persons and caused billions of dollars in property losses. The January 16, 1995 Kobe, Japan earthquake was only a magnitude 6.9 event and yet produced an estimated $200 billion loss. Similar scenarios are possible at any time in Los Angeles, San Francisco, Seattle, and other US urban centers along the Pacific plate boundary. Earthquake physics operates over a wide range of space and time scales, and the development of forecast/forewarning/predictive models together with data mining technologies to understand the dynamics of multiscale systems represents a significant challenge in IT and computational science. Multiscale software technologies developed under this proposal will include general integration capabilities as well as data mining, analysis and visualization tools for correlated, spatially distributed multi-component time series. The present state of simulation and data mining technologies dictate a primary focus on these issues in this first proposal, with less emphasis on the development of data assimilation methods, since in any case, the latter depend on the existence of the former. We will, however, continue to refine our *static* data assimilation methods, while investigating the future requirements for *dynamic* data assimilation using adjoint compiler software such as ADIFOR [8,9].

About the GEM and ACES Collaborations: In recognition of the critical importance of understanding the physics of earthquake fault systems, and the need for developing any kind of forecast or forewarning methodology, both the US and international communities have recently formed teams to focus on the construction of computational simulations of earthquake fault systems that can function as numerical laboratories to investigate the space-time behavior of fault systems. The US group is a collaboration of investigators, the “General Earthquake Model” (GEM) working group chaired by Dr. Andrea Donnellan of the Jet Propulsion Laboratory, Pasadena, California [10]. Complementary to this group is the APEC Cooperation for Earthquake Simulations (ACES), a group of Asia-Pacific economies (Australia, the United States, Japan, China and soon others) that have a formal, US State-department approved collaboration in earthquake simulation [11]. The US GEM group has been in existence for about 4 years, the ACES group for about 3 years. The GEM group meets periodically in the United States in federal agency-funded workshops, while the ACES group has met formally in February, 1999 in Brisbane & Noosa, Australia, and in October 2000 in Tokyo & Hakone Japan. The next ACES workshop is planned for July 29-Aug 3 in Maui, HI. The Japanese have a ~$400 million program to develop the software and hardware technology for earthquake simulations [12]. The GEM group partners actively with the US Southern California Earthquake Center (SCEC), an NSF center focused on data collection and analysis of earthquakes in Southern California [13]. This current proposal seeks funding to continue and accelerate the work of the academic sector of the US GEM program.

About Our Institutional Partners: The Exploration Systems Autonomy section of the Jet Propulsion Laboratory (JPL/ESA), of which Dr. Donnellan is Deputy Section Manager, are full partners in the proposed work, and have a variety of internal and NASA funding to support their efforts (letter). We are also a funded project of the Maui High Performance Computing Center [14]. In addition Dr. Steven Ashby, Director of the Center for Applied Scientific Computing (CASC) at Lawrence Livermore National Laboratory [15], has recently expressed considerable enthusiasm about partnering with the GEM group (letter). CASC has established capabilities in data mining, visualization, Finite Element Methods (FEM), and PSE technologies that can strongly enhance the proposed work. Our JPL/ESA partners add major human capabilities and software resources to the proposed project make successful completion of the work a high likelihood. Should we be able to secure DOE funding for our LLNL/CASC partners, these human resources will be greatly enhanced.

***Computational Considerations and Resources*:** With the assistance of personnel at the Exploration Autonomy Systems (ESA) division at the Jet Propulsion Laboratory, we will optimize the parallel-code implementation of the Monte Carlo-based Virtual_California simulation, both as a Boundary Element Computation (BEC), as well as FEM version. Production of a complementary Virtual-California simulation will be carried out using as a basis the GeoFEM technology recently completed by our Japanese partners in ACES [16]. The Japanese GeoFEM group are part of the ACES working group in computational environments led by Fox. We will also research with our JPL partners development of a fast multipole (FM) version of BEC after we study the results of Tullis on a single fault problem where he is building on standard multipole libraries. Early results suggest that the FM approach outperforms the simple $O(N^2)$ algorithm after 10,000 elements and that Terascale machines will allow million element simulations of fault systems. At present, the BEC code runs on Beowulf cluster configurations using message passing. We are optimizing this code to run on shared memory machines and converting it from F77 to F90/95.

For our computations, we are using resources provided by NCSA (Alliance), in particular the 32 node IBM SMP Squall machine at the Maui High Performance Computing Center (MHPCC), as well as in future the new 260 node, Huinalu SMP IBM Netfinity Linux SuperCluster (Peak ~478 Gflops). In addition, our LLNL/CASC partners will grant access to their institutional computing and visualization resources, which include the Compaq TeraCluster (CTC, 512 processors, Peak ~0.7 TFlop).

Computational MSPSE and Database System: This proposal will use a novel PSE that has been developed to support both education and computing, and will add to it the special capabilities needed by the possibility for earthquake forecasting, namely the initial support for data mining and real streaming of sensor data for simulations at multiple physical and temporal scales. This Virtual Earthquake Analysis Environment (VEAE) will make it possible for groups of distributed collaborators to conduct simulations, analyze distributed data, and publish and share results based on the composition and analysis of these resource objects. The environment will be set up as a MSPSE (Multi-Scale PSE) and will be designed to support the analysis of both observational data and simulations, and later this will be extended to integrated data assimilation. The work will contribute directly to computer science research as well as providing a new operational model for large research teams such as GEM and ACES. A key feature of VEAE is the integrated support of collaboration (sharing) and access and storage of information using XML based meta-objects, which contain just the needed metadata to perform these services. The latter control the active data and simulations/pattern analyses characteristic of this problem area. Through the GEM effort our JPL partners have developed XML Schema to describe various parameters of earthquake faults and input data. As described later the VEAE is back ended by a distributed database (DBMS) system that is especially important in the data mining area and designed to support seamlessly distributed dynamic data. We will develop our earthquake fault databases to focus on paleoseismic, GPS, InSAR, and seismicity data [17]. We also plan to work with communities that have begun to establish data standards, such as the seismic community (effort led by Berkeley), and the International GPS Service. These will be extensions of our XML Schema needed to define the objects.

Data Mining and GeoInformatics: Data mining (GeoInformatics) methods will be refined and further developed to detect space-time patterns of earthquakes using simulations, and then applied to natural earthquake fault systems. This activity, to be based on established methods such as Karhunen-Loeve analysis, Hidden Markov Methods and Ensemble forecasting, will be largely generic in nature, and will be applicable to the analysis of space-time pattern formation and evolution in many kinds of threshold systems [18-20]. The data sets involved are 1) Paleoseismic observations of historic earthquakes whose occurrence and locations are preserved in offset surficial sediments; 2) Earthquake seismicity (origin time, location, magnitude); 3) Surface deformation measured via Global Positioning System (GPS) networks such as the Southern California Integrated GPS Network (SCIGN), and the Bay Area Regional Deformation (BARD) network; and 4) Synthetic Aperature Radar Interferometry (InSAR) observations of surface displacement. Observations of these data types are also planned as part of the Earthscope NSF/GEO/EAR/MRE initiative. The Plate Boundary Observatory (PBO) plans to place more than a thousand GPS, strainmeter, and deformation sensors along the active plate boundary of the western coast of the United States, Mexico and Canada, at a cost in excess of $100 million [17].

Earthquake Physics: The physics of earthquakes is strongly coupled across all scales of space and time. Examples of some of the important spatial scales are shown in the following table. No time scales are shown, since this would require a third axis. Although all scales are important, our current emphasis is on the fault network scale, since this is the scale of most current and planned observational data networks. Because the physics is coupled across all scales, numerical simulations are the only means to provide an integrated, comprehensive understanding of the system dynamics. GEM and ACES scientists feel that a Grand Challenge-scale effort is urgently needed if any sort of forecasting, forewarning, or prediction technology is ever to be developed.

Spatial Scale	Physics	Input from Lower Scale	Output to Upper Scale	Comp. Methods	Research Status
Atomic: 10^{-10}m -1 μm	Quantum. Disordered system	Fundamental atomic constants	Cohesive potentials	Quantum DFT, MC	Computational chemistry, Not proposed here
Grain size 1 μm - 1 cm	Contact interactions, planar fault elastic walls	Cohesive potential across grains	Effective viscosity, LG effective constants	MD	Not proposed, but we use input from ACES partner research
Fault zone 1 cm - 100m	Fluidized viscous gouge, elastic walls & interactions, strong correlations	Effective viscosity, LG effective constants	Effective friction laws, e.g., rate & state, stick-slip, leaky stress, elastic constants μ, λ, effective LG constants	FD, FEM, CA, BEM, Inertial solvers	Not proposed, but we use effective parameters (μ, λ, α) from literature, ACES, and our other research.
Fault system 100m - 10 km	Coarse-grained planar faults, effective friction, strong correlations	Effective friction laws, e.g., rate & state, stick-slip, leaky stress, effective elastic constants μ, λ, α, effective LG constants	Effective elastic moduli μ, λ, effective coefficients of friction & stress leakage (μ_S,μ_K,α), effective LG constants	CA, BEM, FEM, Quasistatic solvers	Basis of our statistical mechanics approach. We use results from literature and our other funded research
Fault networks 10 km - 1000 km	Geometic fault complexity. Deep viscoelastic relaxation, Static-kinetic friction, strong correlations	Effective elastic moduli μ, λ, effective coefficients of friction & stress leakage (μ_S,μ_K,α), effective LG constants	Effective viscosity spectrum, Effective viscoelastic modulus spectrum	CA, BEM, GeoFEM	Work proposed here is primarily on this scale.
Tectonic Plate Boundary	Viscoelastic flow on very long time scales, kinematics of plate motion at fault velocities V	Effective viscosity spectrum, Effective viscoelastic modulus spectrum	No larger scale of interest	GeoFEM	Models on this scale give us fault loading velocities V

Table Key: FEM - Finite Element Method; GeoFEM - Japanese Geo FEM; MC - Monte Carlo; DFT - Density Functional Theory; FD - Finite Difference; MD - Molecular Dynamics; PD - Particle Dynamics; LG - Landau Ginzburg; CA - Cellular Automata; BEM - Boundary Element Method

Scientific and Computational Objectives: Our proposed work will:

Improve Pattern Analysis Techniques for analyzing multi-scale observational data and the results of simulations using methods from statistical physics, together with theoretical investigations of pattern-based data mining.

Improve and Optimize Earthquake Simulation Technology by introducing Fast Multipole Methods and other acceleration algorithm techniques, as well as prototype the use of large SMP machines for simulation code development and use.

Construct a Multi-Scale Computational Problem Solving Environment for pattern-based and ensemble forecasting, and integrating data analysis and modeling with pattern dynamics analyses and other data mining methods.

Link Multi-scale Database Management System with software technologies for shared collaboration, and interfacing with non-specialists and others. In the best case, this environment may enable the development of an adaptive observation technology, particularly relevant at the time of a major earthquake.

Improve Our Current Static Data Assimilation Methods, and Define the Requirements for the use of adjoint compilers developed at JPL and elsewhere, to improve forecast power and performance of simulations, and to accelerate the introduction of ensemble forecasting.

Data Mining & GeoInformatics: Using simulations and real data, we will focus our efforts on three different but complementary approaches to earthquake pattern recognition. The first is a new pattern dynamics method (Phase Dynamical Probability Change, or PDPC), which uses the mathematics of pure phase dynamical systems to describe space-time correlations of seismic activity [18-21]. The observable earthquake activity represents a "wave function," and the underlying dynamics are associated with "hidden variables." This method has the advantage that relative probabilities can be readily defined given the mathematical phase dynamics framework. We show the results of applying this pattern dynamics type data analysis method to instrumental earthquake data recorded by regional seismic networks since 1932 (Figure 1) [18,21]. The data is in the form of a record of approximately 60,000 earthquakes of magnitude greater than 3.0 occurring in California. In Figure 1, the colored regions locate sites of enhanced Log (probability) for earthquakes larger than magnitude M> 5 that developed over the years 1989-1999. Some of these regions are associated with the actual events of M>5 that occurred during 1989-1999, but other colored regions are as yet unoccupied by large events. These latter regions represent forecasts for future major events over approximately the next 10 years. Retrospective studies of this method using random catalogs and statistical likelihood ratio tests indicate that the pattern dynamics method has considerable forecast skill [21].

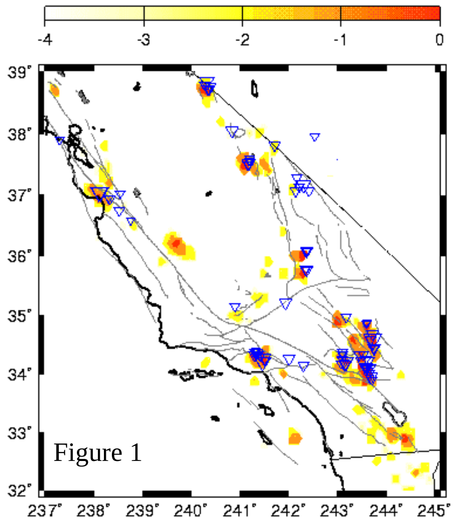


Figure 1

The second pattern recognition approach is based around the use of Hidden Markov models (HMM) [22]. In the HMM framework, the observable data is assumed to have been generated by an underlying stochastic process, which is at any time in one of a set of distinct states. Each state is described by a probability distribution of its observable output, and by the probability of the system being in the same state, or in each of the other states, at the next point in time. Given the observations and a few selected model parameters ("feature vector") it is possible to objectively calculate a model for the system that generated the data, as well as to interpret the observed measurements in terms of that model. Furthermore, because the model includes the probability of each state transitioning to each other state at the next time step, it possesses predictive power (i.e., the state of the model at the last observed point in time is known, so the most probable state and output at the next point in time are also known).

Using HMM-based methods to explore data generated by complex physical systems is particularly challenging, as the resulting model must be both consistently reproducible and scientifically meaningful. However recent advances, including large-scale numerical methods, simulated and deterministic annealing [23], and automated determination of the number of states [24], can be applied to this problem to make HMM-based analysis a valuable tool for analysis of geophysical and other scientific data. We have carried out a preliminary analysis using this method applied to a record of approximately 2,000 earthquakes with magnitude greater than 4.0 that have occurred in Southern California. We find two important classes of earthquakes, one of which includes several major events, including the Hector Mine and Landers earthquakes, and the other of which is a class of earthquakes having a relatively large magnitude and a long time until next event. The transition probability P_T for class 14 $\Rightarrow$ class 16 is $P_T \sim .12$, significantly larger than from class 14 to any other class. This result indicates that if an event of class 14 is found, the next event is most likely to belong to class 16, the class of large earthquakes. While the relationship between these two classes has not yet been fully investigated, this example does demonstrate the potential of this kind of analysis.

Proposed Work: Proposed work for data mining issues relating to earthquake pattern detection and interpretation will be our primary algorithm-development and computational implementation activities. These activities will be in three distinct areas. 1) We will extend our PDPC to smaller spatial scales and shorter time scales. Currently, the method is optimized to search for space-time correlations indicating possible future activity on spatial and temporal scales corresponding to earthquakes greater than magnitude 6. While theoretical investigation of the PDPC method to promote improved physical understanding continues, it is nevertheless straightforward to extend these methods to examine smaller and more frequent classes of events, so that statistical evaluation of the method is accelerated. To optimize this process, we also plan to integrate the PDPC within the VEAE described above. 2) We will extend the HMM methods to include more possible features and classes. As part of this effort, we plan to fuse the HMM method with the Karhunen-Loeve seismic activity eigenvector methods we have developed previously, thereby creating a hybrid method in which the HMM feature vector and catalog of possible classes is defined via eigenpatterns rather than by less objective means. The integration of these two methods within the VEAE will in effect define a Hidden Markov-type mode shaping approach to pattern analysis and recognition. 3) Although our current simulation technology does not at the moment employ dynamic data assimilation methods associated with ADIFOR-type adjoint compilers [8,9], we can nevertheless prototype a somewhat more primitive, but still probably effective form of ensemble forecasting. The key to this method is the recognition [25,26,27] that the mean field nature of the simulations implies that all possible configurations of large events are visited by a given simulation eventually (discussed below). Given the static data assimilation methods we currently employ [18,28,29], we have demonstrated that our simulation technology produces quite realistic catalogs of simulated events. Using the data base methods discussed previously and below, we plan to create catalogs of simulation events for the major faults in southern California, then datamine for sequences that are maximally similar to the known recent history of events in Southern California. The earthquakes following these specific sequences then represent a projection, or forecast of future activity, on the real fault system, subject to the usual caveats concerning model errors. The entire ensemble of these sequences can then be used as a way of evaluating the probability of given future events in California. To realize this program in an effective way, both simulations and data mining methods must be integrated with the VEAE.

Computational Environment: The VEAE is built around modern distributed object technology communicating with a dynamic publish/subscribe message system. Objects include computers, users, reports, presentations and the information needed to specify input and output of simulations initiated by any of geographically distributed researchers. Field data and information streaming from sensors are also objects. The VEAE will use software infrastructure initially designed to support education and a computing portal Gateway for materials science. This proposal will adapt the existing environment and focus on the special needs of complex systems analysis with special attention to multi-scale problems.

The base infrastructure of the VEAE has several novel features that have designed to address issues coming from both previous research [30] and a detailed analysis [31] of major commercial tools in the collaboration and object management area. We will exploit existing systems (Garnet, Gateway: [31,32]) that provide a web interface to access and to manipulate the earthquake science objects, and to allow this to be done collaboratively. This capability is formulated as the Garnet Collaborative Portal (GCP: [33]), which uses an integrated distributed object framework to specify all the needed object properties including both their rendering and their collaborative features. GCP Inclusion of rendering information allows one to customize to different clients and to effectively build collaborative environments integrating hand held and desktop clients. We assume that we are building on a computational grid infrastructure and so can layer our high level services on top of the capabilities under development by projects such as the NEES Grid being developed for earthquake engineering and for experimental physics, the Particle Physics Data Grid and GryPhyN [34-36]. Users, Computers, Software applications, Sessions, and all forms of information (from field data to simulated events to recording of audio/video conferences) are all objects, which can be accessed from GCP. These objects will all be self-defining; namely make explicit all the necessary metadata to enable GCP to perform needed functions such as searching, accessing, unpacking, rendering, sharing, specifying of parameters, and streaming data in and out of them. This metadata is defined using a carefully designed XML schema GXOS and exploiting the new RDF framework from W3C [37]. The XML meta-objects that define the GCP point to the

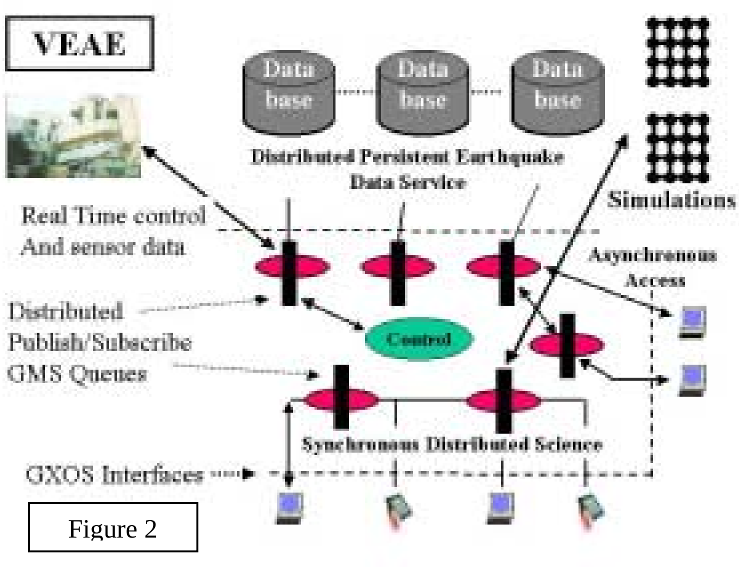


Figure 2

location of the object they define and can initiate computations and data transfers on them.

VEAE software is largely written in Java (using Enterprise Javabeans in the middle tier) but Java/XML is only the execution object model of the meta-objects; one can load persistently stored meta-objects or control target base objects formed by flat files, databases, CORBA, .net (SOAP) or any distributed object system to which we can build a Java gateway. Our successful Gateway computational portal has used this strategy already; here all object interfaces are defined in XML but CORBA access is generated dynamically. Further this system also only uses meta-objects in the middle tier and invokes programs and files using classic HPCC technology such as MPI. This strategy ensures we combine the advantages of highly functional commodity technologies and high performance HPCC technologies.

GCP uses the shared event model of collaboration where these events use the same base XML schema as the meta-objects describing the entities in the system. The uniform treatment of events and meta-objects enables us to use a simple universal persistency model gotten by a database client (Figure 2 above) subscribing as a client to all collaborative applications. Integration of synchronous and asynchronous collaboration is achieved by the use of the same publish/subscribe mechanism to support both modes. Currently we use JMS (Java Message Service) to provide publish/subscribe services for events in our prototype GCP but have already found serious limitations that we will address in a new GMS (Garnet Message service). This uses GXOS objects to specify both topics and user profiles and can be considered as a candidate for a Grid Message service. The control module of a collaborative session shown in Figure 2 is also implemented as a GMS client. Innovative collaborative features in GCP include simultaneous support of desktop and hand held devices and collaborative viewers for shared display, and a set of important export formats – PDF SVG, Java3D and HTML. We are also designing a control script based on the W3C Resource Description framework (RDF) to describe where resources are located and to support object services such as create and copy. This could be viewed as developing an XML based “operating system” with goals related to those of Berkeley’s Ninja system[38] and the W3C Semantic Web[39].

There is a major effort to understand computational portals within the Grid Computational Environment working group of which Fox is a co-lead. GCP can be considered as a computational portal to which we have added a publish/subscribe based collaborative service and for which a uniform set of object schema are used to define sharing, security, rendering and access properties. This Grid forum [40] activity is working with many different portals and describing them with a common template [41-42]. We will continue to work with this international group so as to ensure that we feed important lessons such as GMS into the Grid architecture and that we bring to the earthquake field the best ideas for distributed computational environments.

Proposed Work: The GCP prototype available in May 2001 will only implement basic versions of these ideas and both the general framework and the extensions needed for GEM applications imply major research and implementation challenges. Research areas include the Grid Message Service as well as the Grid server architecture. We will also research the structure of the scalable management of meta-objects in an efficient effective fashion. Whether it is a reasonable model to separate object and meta-objects will be an important lesson from this research. Our use of RDF as the scripting language of the overall system will challenge this relatively simple meta-data model and perhaps point the way to improvements. Integrating the event models between the different event models and synchronous and asynchronous collaboration modes is another hard area. In this proposal we will focus on the specialization of the object model to GEM – an effort that will test the design principles for GXOS as well as forcing a careful analysis of the computational process of GEM. We will extend the initial work of JPL on XML schema for earthquake faults and GEM related sensor information and specify the objects needed in the VEAE as extensions of GXOS.

GCP has many features of value to the field such as real time collaborative analysis of sensor data, and the use of the system by GEM researchers will lead to many lessons and consequent advances throughout the system. However our VEAE research focus will be on using the collaborative portal to support multi-scale analysis in a MSPSE – Multi-scale problem solving environment. Examining the table we suggest that a MSPSE should allow

a) Specification of all aspects of problem (software, data, results, associated resources) in a hierarchical fashion that reflects the science scaling.

b) Linkage of simulations and data at different scales where the fine scale work is abstracted as a law (such as a chemical potential or friction law), which can be used as input to a larger physical size but coarser scale computation.

c) Hybrid simulations where multiple scales are simultaneously present in a given run with different space or time segments modeled at the abstraction level needed by the physical situation.

These are just a first list of characteristics and one research issue will be to refine and extend these characteristics of a MSPSE. Any real case involves aspects of all these characteristics. As a model, a strict hierarchy is somewhat idealized, and one needs to be able to specify multiple hierarchies. For example, earthquake engineering is often linked to GEM like simulations but this field would be specified separately from table above. The hierarchical structure of GXOS naturally supports a) with multiple hierarchies allowed, and further one has the ability to use full XLink/Pointer linkages to "jump" across object hierarchies. Characteristic c) can be addressed by composite tasks (specified by generalized dataflow) in the Gateway portal component of GCP. Natural ways of supporting b) and quantifying how best to realize a) and c) in a PSE will be central research issues. We expect to use XSIL [43] from Caltech to specify the scientific parameters needed to specify abstractions in a). We already have experience in this from Gateway.

Data Base Management System: Database and information management facilities are required to store, access, integrate, and mine information from the various collections and simulations associated with this project. The information sources are heterogeneous in nature, in the organization of their content as well as the software used to manage them. Our research approach here is to devise, develop, and utilize techniques and mechanisms to support the importation of assembled complex information objects into the VEAE MSPSE. We propose to assemble and transmit information to/from a MSPSE in XML, in response to a request supplied via MyXoS from either a simulation, a pattern analysis or user. This XML will extend the existing GXOS Schema so that for MyXoS, any data base query is a GMS message, which can be uniformly handled by MyXoS. Critical research issues include the integration of the database middleware engine with GMS event service and the efficient generation of the metadata used in the meta-objects controlling the collaboration and rendering of objects in GCP. Both the databases and the MyXoS servers are dynamic and distributed. Objects whether pre-existent or real time sensor streams and simulation results publish their identity on queues managed by

MyXoS. The database system must archive both the GMS events and these incoming data streams.

Storage and access of this static and dynamic information will be managed by several systems, which are associated with the specific collections/databases. New databases(s) will be managed by an extended relational database management system. To support the integration and fusion of data from multiple databases and heterogeneous sources, we propose to employ ontology-enhanced wrapper-based techniques [44-49]. In this approach, data sources are made to appear as virtual databases, despite their heterogeneous structure, and despite the fact that some sources are not full databases. To specifically deal with the definitional diversity (semantic heterogeneity) among the source databases/collections, we employ techniques based upon a domain knowledge base (ontology), and a technique termed "meta-data implantation and stepwise evolution" [44]. The latter is a partial database integration scheme in which remote and local (meta-)data are integrated in a stepwise manner over time. The meta-data implantation and stepwise evolution techniques can be used to inter-relate database elements in different databases, and to resolve conflicts on the structure and semantics of database elements (classes, attributes, and individual instances). The approach employs a semantically rich canonical data model, and an incremental integration and semantic heterogeneity resolution scheme. Relationships between local and remote information units are determined whenever enough knowledge about their semantics is acquired. The folding problem, folding remote meta-data (conceptual schema) into the local meta-data to create a common platform through which information sharing and exchange becomes possible, is solved by implanting remote database elements into the local database, a process that imports remote database elements into the local database environment, hypothesizes the relevance of local and remote classes, and customizes the organization of remote meta-data.

Proposed Work: We specifically propose to support key kinds of information interconnection, explained here as connections between an importing collection and an exporting collection. These interconnections are communicated as GMS events and will support the MyXoS meta-objects: 1) The **Direct Link** method imports only the references to the sources of imported information and each subsequent access will result in fetching data from their source databases. It is most suitable for frequently changed source. 2) The **Copy** method duplicates all the data of interest on the importer database. It has the best performance because copies of data are stored locally at the importer side and there is no need to retrieve data through network. 3) The **Caching** sharing mechanism is a compromise. The importers keep both references to data sources as well as local copies. It has better access performance than direct-link since the importers can use local copies, and but it introduces additional overhead to the exporters. 4) **Time-bounded copy** is a special case of the caching method. It targets those data which are updated periodically, but the time-bounded sharing mechanism is only suitable for those data which are updated regularly. 5) The **Remote-query execution** method is designed to address the occasions when the set of data is so large that it is more practical to execute a query request on the exporter's side rather than bring the whole data to the importer's side.

The sharing primitives/tools to be developed in this research will support a scientist MSPSE, allowing a researcher to maintain a personal "view", in particular tracking relevant subsets of information and their interconnections. The sharing primitives/tools and ontological conversions (conversions from one concept / inter-relationship / terminology framework to another) can support inter-database referencing at various levels. Developing and deploying such techniques will allow referencing database subsets across database boundaries.

One of the key results of this work will be support for establishing interconnections and relationships among data at various levels of abstraction. Thus, a principal anticipated result of this work is support for the study of scientific data, models, simulations, and analysis from the atomic to tectonic plate boundary scales.

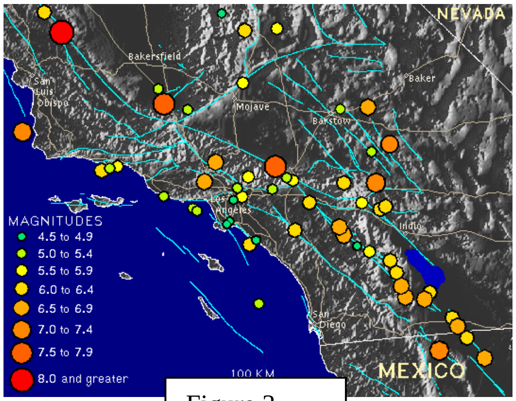


Figure 3

Numerical Earthquake Simulations -- Algorithms: General methods for carrying out the network simulations have been discussed in refs. [18,19,28,29,50,51]. Here we provide a brief summary of one particular set of algorithms that define the Virtual_California simulation, as an illustration of how an entire class of earthquake simulation methods is realized. As shown in [25-27,59], these meanfield models are Abelian, with the final slip depending only on the initial state of stress and the difference in the friction coefficients and not on the order that segmens slip, so details of the friction laws are neither observable nor relevant. It is important to note that many of the functionalities of the APIs that have been used to code these algorithms in the past will be greatly enhanced within the context of the proposed MSPSE.

Elastic Interactions: To begin, one defines a fault geometry in an elastic medium based on a map of faults in a tectonic region, computes the elastic stress stress transfer coefficients, or interaction Greens function tensors $T_{ij}^{kl}(\mathbf{x}-\mathbf{x'})$, assigns frictional properties to each fault, then drives the system via the slip deficit (defined below). The elastic interactions produce mean field dynamics and strong correlations in the simulations [25-27]. We focus here on the major horizontally-slipping strike-slip (horizontal motion) faults in southern California that produce the most frequent and largest magnitude events. We used the tabulation of strike slip faults and fault properties as published in ref [52]. All major faults in southern California, together with the major historic earthquakes, are shown in Figure 3. Figure 4 shows our model fault network.

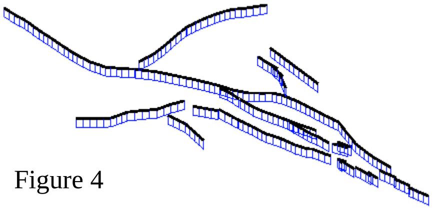


Figure 4

Friction: Several friction laws are described in the literature [51,53], including Coulomb failure , slip-dependent or velocity-dependent friction, and rate-and-state. Here we use a parameterization of recent laboratory friction experiments [51]. Upon loading, a failure threshold $\sigma^F(V)$ is reached at which sudden slip occurs and the stress rapidly decreases to the level of a residual stress $\sigma^R(V)$, where V is load point velocity. Stable precursory slip is observed whose velocity increases with stress level, reaching a a few percent of the load point velocity prior to failure at $\sigma = \sigma^F(V)$. The simplest form of the friction equations describing these experiments can be obtained from space-time coarse-graining procedures applied to simple planar fault models [29]:

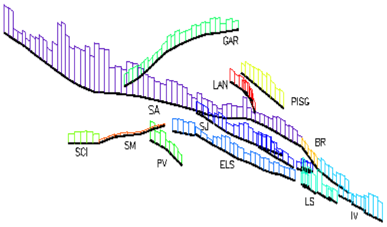
$$\frac{ds}{dt} = \frac{\Delta\sigma}{K}\left\{\alpha + \delta\left(t - t_F\right)\right\} \qquad \text{(Friction Stress)} \qquad (1)$$

$$\sigma = K\left(Vt - s\right) \qquad \text{(Elastic Load Stress)} \qquad (2)$$

where $s(\mathbf{x},t)$ is slip at position $\mathbf{x}$ and time t, $\sigma(\mathbf{x},t)$ is shear stress, K is a elastic stiffness (change in stress per unit slip), and stress drop $\Delta\sigma = \sigma - \sigma^R(V)$. For laboratory experiments, K is the contact layer stiffness, and for simulations, represents the stiffness of a coarse-grained element of the fault of scale size L. α represents the ratio of the effective stiffness modulus KL to a surface viscosity η, $\alpha = KL / \eta$. $\delta()$ is the Dirac delta, and t_F is any time at which $\sigma(\mathbf{x},t_F) = \sigma^F(V)$. The quantity $(s - Vt)$ is the slip deficit referred to above. Both σ^F and σ^R can be parameterized as functions of the normal stress χ by means of coefficients of static μ_S and ("effective") kinetic μ_K coefficients of friction, $\sigma^F = \mu_S \chi$, $\sigma^R = \mu_K \chi$. The latter relation implicitly assumes that dynamical overshoot or undershoot during sliding is approximately constant [53].

Static Data Assimilation: Earthquake data obtained from the historical record as well as geological field studies represent the primary physical signatures of frictional properties on faults. The timing, magnitude and complexity of these historical events are a direct reflection of the values of the frictional parameters: α, σ^F, σ^R, all of which should be regarded as scale-dependent functions. For historical earthquakes, there can be considerable uncertainty about where the event was located [54]. Modern studies [54-56] indicate that slip is often distributed regionally over a number of faults and sub-faults. Therefore our technique assigns a weighted average of the slip for given historic events to *all* of the faults in the system, with most of the slip assigned to faults near the location of maximum ground shaking. Since the seismic moment is the torque associated with one of the moment tensor stress-traction double couples, it is most reasonable to use the (inverse cube power of distance r) law that describes the decay of stress with distance [28]. Comparisons with data indicate that this method yields average recurrence intervals similar to those found in nature.

Figure 5



Step 1: Assignment of Moment Rates. All historical events in southern California since 1812 are used (ref [52]). The seismic moment, a measure of slip during an earthquake is $M_o(t_j) = \mu \left\langle s(t_j)\right\rangle A$, where μ is shear modulus, $\langle s(t_j)\rangle$ is average slip at time t_j, and A is fault area. For each of the fault segments, the contribution of moment release rate from the jth historical earthquake $dM_o(t_j)/dt$ to the rate on the ith fault segment, $dm_i/dt = \Gamma \left[\Sigma_j \{dM_o(t_j)/dt\} r_{ij}^{-3} \right] / \Sigma_j r_{ij}^{-3}$, where $r_{ij} = |\mathbf{x}_i - \mathbf{x}_j|$ is the distance between the event at location $\mathbf{x}_j$ and time t_j, and the fault segment at $\mathbf{x}_i$. The factor Γ accounts for the limited period of historical data available compared to the length of the earthquake cycle, and is determined by matching the total regional moment rate, $\Sigma_i\, dm_i/dt$, to the observed current regional moment rate. We find $\Gamma \approx .44$.

Step 2: Determination of Friction Coefficients. For a compact fault, the average slip in terms of stress drop $\Delta\sigma$ is [57,58]

$$\langle s\rangle = \frac{f\ \Delta\sigma\sqrt{A}}{\mu}$$

where f is a dimensionless shape factor. Standard assumptions of $f \sim 1$, $\Delta\sigma$

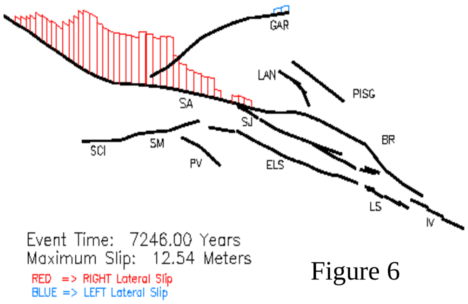


Figure 6

~ 5 x 10^6 Pa, μ ~ 3 x 10^{10} Pa yield reasonable slip values. The average slip is converted to a difference between static and kinetic friction, $(\mu_S-\mu_K)_i$ for the i^th^ fault segment via the relation: $(\mu_s - \mu_k)_i \approx \frac{m_i}{f\, A^{3/2}\, \chi_i}$ obtained by combining the two previous relations , and $\Delta\sigma \approx \sigma^F - \sigma^R$. To compute $(\mu_S - \mu_K)_i$, a typical value of χ_i for each segment is computed from the average gravitationally-induced compressive stress. Since the stochastic nature of the dynamics depends only on the differences $(\mu_S - \mu_K)_i$ [18,19,28,29], we set μ_k = .001.

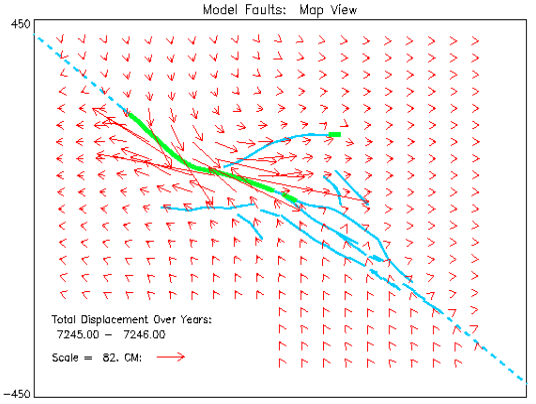


Figure 7

Step 3. Aseismic Slip Factor: Earthquake faults are characterized by varying amounts of aseismic creep (slow slip generating no elastic waves) that arises from the "stress leakage" factor α. The average fraction of slip on each fault that is aseismic is equal to α_i./2, and has also been tabulated for southern California faults in ref [52].

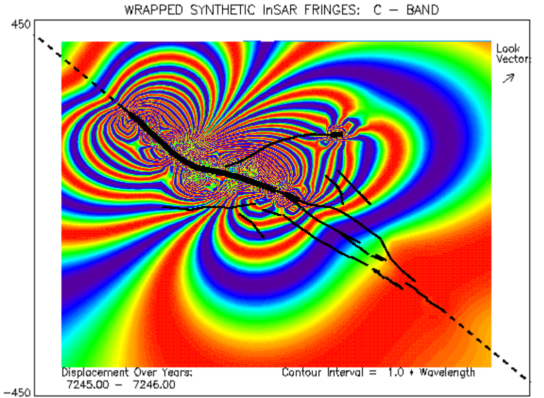
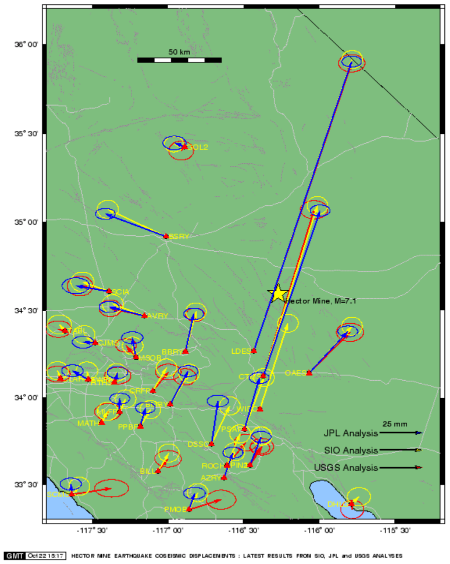


Figure 8

Results: Figures (5-9) show examples of the simulation results we have obtained. Figure 5 is a plot of μ_s - μ_k for all fault segments. Figure 6 is an example of the slip in a simulated San Andreas earthquake. Figure 7 is a plot of horizontal GPS-type surface displacements from the same event, and Figure 8 is a plot of surface displacments that would be observed via InSAR observations. Figures 9 and 10, both produced by JPL, are examples of GPS and InSAR data obtained for the October 16, 1999 Hector Mine earthquake.

Figure 9



Friction Models and Acceleration Algorithms: Our group has pioneered the use of a variety of friction models in earthquake simulations. There is no known friction law obtained from laboratory experiments that is appropriate to high-velocity sliding, so details of the deterministic dynamics of sliding for any friction law will be approximate. However, the Abelian nature of our meanfield models [59] implies that the final state of slip can be computed with more confidence. In one of our friction models [50], both a critical scale L_c for earthquake nucleation, and a rupture healing scale L_H are defined (L_H >>

L_c). This friction law is both velocity- and slip-dependent, as well as propagating stress information at the speed of P-waves. Numerical experiments have revealed little sensitivity of the final state to the presence or absence or waves, as would be expected of Abelian (mean field) models. Another friction law that our group has used extensively is Runge-Kutta integration of rate-and-state friction, implemented in planar fault models of the Parkfield segment of the San Andreas fault [51, 60]. The formulation used is a composite of the "slowness" and "slip" laws, with the model behaving as the slip law at high sliding speeds, and as the slowness law at lower speeds [61]. The Parkfield simulations have also included radiation damping terms, which serves to introduce a time scale into the sliding dynamics. As part of the GEM effort, we are also developing an application of the Fast Multipole method (FM) that will accelerate future simulations to scale approximately as N log N in computational time, rather than N^2, where N is the number of fault segments. FM methods were originally developed for problems in astrophysical N-body dynamics and in computational fluid dynamics [62,63]. Members of our group have considerable experience in FM methods.

Proposed Work: The primary work we propose is to scale our existing simulation technology to significantly larger system sizes using the computational resources described earlier, and to integrate these simulations with the MSPSE and the DBMS described above. In particular, we will optimize our simulation algorithms within the object-oriented client/server and P2P-type environment defined by the VEAE and GCP technology. Particular issues for implementation will involve both CA and Runge-Kutta type time-marching, as well as FM methods for full three dimensional simulations. We will also integrate these results with the visualization and rendering technology included in the VEAE. Implementation of our simulations will be compatible with the developing Grid standards and architecture.

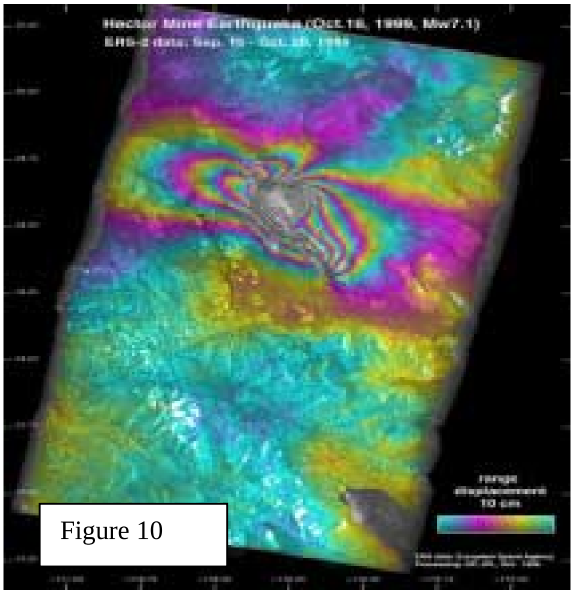


Figure 10

Knowledge Transfer, Education and Public Outreach: Hazard mitigation requires dissemination of knowledge about our current understanding of natural hazards. In addition to providing information about earthquake simulations to other professionals and to the public, we believe that an important role of outreach is to educate the broader public on the scientific method used in natural hazards research, about the evaluation of natural and human-influenced hazards, and about the role of IT in science and engineering. The most effective way to accomplish these goals is to extend outreach both to professionals, as well as to current and future K-12 teachers. This latter community of educators reaches a much broader spectrum of the public than any single outreach program can, thus there is great value added by orienting outreach towards K-12 teachers.

Proposed Work: 1) We will participate in the training of future high-school science teachers by providing summer research opportunities in earthquake science. The University of California, Davis, is developing a new Bachelor of Science major in

Natural Sciences, designed to prepare students to meet the standards of California's Science Teaching Credential. This program, which will be administered by the Geology Department, is currently undergoing formal review at the campus level and we expect it to be approved. We plan to involve several students from this program (and similar programs at the other campuses) in summer research internships working with the primary investigators of this program. By actively participating in research, students will gain a greater understanding of the scientific method to take to their future classrooms. Students will receive a small stipend to cover their living and travel expenses while they are interns. If the program is successful, we will seek additional funding from state sources to expand it to current science teachers.

2). Develop a web interface for the General Earthquake Model designed for K-12 teachers, based on a teaching "portal" modeled on the collaborative portal GCP that is the basis of the proposed research program. The teaching portal will provide K-12 teachers with access to information about both the science and IT aspects of earthquake research. The content will be developed primarily by an education and outreach professional at the University of Colorado. The system will allow distance delivery of seminars and classes.

3) We will transfer knowledge to professional colleagues, primarily through tutorial workshops organized by ourselves and the Southern California Earthquake Center (the next such meeting is July 29-Aug 3 at MHPCC, Maui, HI), professional meetings, web-based communications, and peer-reviewed papers. We are using all these methods with considerable success.

Roles of the Investigators & Management Plan: Dr. John Rundle is the PI of this proposal and will oversee all aspects of the work. The team members have worked extensively and effectively together in the past. Managing such a large team can be difficult, so we plan to hold regular workshops in which all of the team members can interact and plan work. Primary team members will be responsible for developing and completing different objectives of the work as outlined in the following table. Steve Ashby and LLNL/CASC personnel will provide advice and interface where appropriate.

Objective	**Team Members**	**Task, Background, Skills**
Pattern Analysis & Data Mining	John Rundle Dennis McLeod William Klein Robert Granat Andrea Donnellan	Phase Dynamics Informatics Phase Dynamics, Statistical Physics Hidden Markov Models Data Inversions, Hybrid Methods
Computational Environment	Geoffrey Fox Dennis McLeod	Object Brokers, Environments, Portals Storing, Accessing, Inter-relating Information
Database Management	Dennis McLeod Geoffrey Fox Andrea Donnellan	Database Middleware, XML Data Definition GCP & XML Objects Data Definition & Management
Simulation Codes & Algorithms	John Rundle Terry Tullis Louise Kellogg Steve Ward John Lou William Klein	Simulations & Static Data Assimilation Friction, Fast Multipoles, Simulations Finite Element Methods, Simulations Simulations Fast Multipoles, Simulations Strongly Correlated Systems, Simulations
Data Assimilation & Requirements	John Rundle John Lou	Static Data Assimilation Dynamic Data Assimilation, Model Steering
Outreach	Louise Kellogg Marie Rundle	Outreach to Professionals K-12 Science Content