EarthCube Design White Paper: Focus on Forecasts

Philip Maechling
Information Technology Architect
Southern California Earthquake Center / University of Southern California
3651 Trousdale Parkway, Los Angeles, CA 90815

1) A Vision for EarthCube: What is envisioned as the scope of EarthCube? How will it transform geosciences research? What functionality will EarthCube provide to the Geosciences community?

As a scientific initiative, our vision is that EarthCube produces an organizational and scientific process that helps researchers improve geoscientific forecasts. Our vision is that geoscientific forecasts improvements drive EarthCube scientific, organizational, and cyberinfrastructure development.

EarthCube should focus on forecasts for multiple reasons. Geoscientific forecast improvements emerge from a broad range of geoscientific specialties and no specialties are excluded. Distribution of EarthCube resources towards forecast improvements reduces concern of favoritism towards any specific group. The geosciences produce numerous broad-impact forecasts. Geoscientific forecasts demand interdisciplinary inter-workability. Geoscientific forecast development, operations, and evaluations are exceptionally demanding of computational resources and data management. Geoscientific forecasts are critical interfaces between the geosciences and other domains. Forecast improvements will drive improved communications between geosciences and external groups. Geoscientific forecasts have economic value offering economic sustainability.

As a technical development, our vision is that EarthCube develops a modular, broadly usable, forecast evaluation system that helps research groups improve the automation, reproducibility, and rigor of forecast evaluation. Rigorous, repeatable, well-defined forecast evaluations will transform geosciences. EarthCube system development will focus on developing the tools needed to test and evaluate a geoscientific forecast. The key technical elements needed will include (1) organization of observational and forecast data, (2) access to computational and storage resources, and (3) software automation of forecast and evaluation processing. The EarthCube system will automate evaluation of geoscientific forecasts and their constituent parts, supporting retrospective forecast evaluation for forecast development and prospective forecast evaluation for forecast operations.

The EarthCube system can be used by any geoscientific specialty to evaluate a forecast or geoscientific model. The EarthCube system enables forecasters to communicate their results in convincing way to forecast users. EarthCube becomes a standard tool for establishing the utility of a geoscientific forecast. Specialized EarthCube systems can be combined to enable larger-scale and interdisciplinary forecasts.

As a sustainable scientific activity, our vision is that EarthCube becomes an interface between geoscientific organizations and forecast users. EarthCube, as liaison organization, should identify and communicate with users of geoscientific forecasts and identify forecasts improvements with sufficient economic value to support ongoing EarthCube operations and developments. EarthCube, as a scientific organization, identifies shared observational and computational needs, and establishes forecast standards and evaluation criteria. Users of economically valuable forecasts support EarthCube cyberinfrastructure development. In this vision, NSF supports EarthCube only until the EarthCube scientific community shows that forecast improvements have economic value, at which time, EarthCube is supported by geoscientific forecast users.

2) A Community-Based Governance model: The community structures necessary to acquire current and future user input/requirements, to respond to changing data and science needs, to adapt and adopt new technologies, to coordinate components and facilities, to foster partnerships and community participation.

EarthCube community-based governance model should establish three groups: (1) economic, (2) scientific, and (3) technical.

The EarthCube economic group is responsible for identifying economically valuable and scientifically improvable forecasts. This group will identify forecasts with existing commercial or governmental markets, such as wind, wave, water flow forecast, and ground motion forecasts. There are many existing well defined, commonly used, often regulatory-controlled, geoscientific forecast types. The economic group's task is not to define new types of forecasts. Rather, it is responsible for identifying forecast users that will provide economic support for forecast improvement. The economic group is responsible for developing the sustainability plans for EarthCube. The economic group must identify specific forecasts improvements for which there is economic support. These forecasts then become the development case for EarthCube CI.

The EarthCube scientific group is responsible for developing forecast evaluation procedures. The scientific group reviews the geoscientific forecasts, establishes standards for comparing forecasts, identifies forecast elements that need to be evaluated, establishes standard methods for expressing uncertainties in forecasts, and defines evaluation criteria. The scientific group is responsible for defining the forecasts, and for defining the processing to be done within an EarthCube forecast evaluation system. EarthCube as an organization provides the scientific guidance for both the economic cases, and guidance for researchers that seek to put their forecast under evaluation.

The EarthCube technical group is responsible for implementing the computational and data management infrastructure needed to rigorously evaluate geoscientific forecasts. The technical group establishes the operational concepts and principles for forecast evaluation. The technical group is responsible for prototype implementations and production of open-source distributions. The technical group is responsible for identifying broadly useful interfaces to observational data, to computational resources,

and to simulation results. It is also responsible for establishing interfaces between EarthCube installations and aggregating EarthCube systems together.

3) The Conceptual CI Architecture: The architecture necessary to provide the services of EarthCube, to integrate advanced information technologies that facilitate access to distributed resources such as computational tools and services, instruments, data, and people.

Our EarthCube CI architecture describes basic system capabilities and introduces general operational principles. These principles are then used in detailed decision making over the lifetime of the program. Basing the architecture on principles, rather than on specific technologies, establishes value across multiple disciplines because existing technology can often be applied in ways consistent with these principles.

In EarthCube, a geoscientific forecast evaluation system is the unit of work. Automated forecast testing requires observational data management, earth structural model development, physics-based forecast model development, simulation data management, standardization of forecast types, and consensus on forecast evaluation criteria. This is precisely the work needed to produce scientific results that are more broadly useful, interoperable and inter-workable.

Our EarthCube architecture is modular. Forecasts require widely varying levels of evaluation. This EarthCube architecture represents a generic small-scale forecast evaluation system. EarthCube achieves scale and support for complexity by assembling multiple evaluation systems together through their external interfaces. An overview of EarthCube architecture with significant internal data stores and external interfaces is shown in **Figure 1**.

EarthCube CI architecture is based on essential capabilities rather than specific technologies. Many existing technologies may be useful within an EarthCube installation. EarthCube should be based on establishing principles and example implementation that supports the common needs of evaluation systems.

Within EarthCube, data and software should be inseparable, and should managed together. No data will be registered into a system without software to read it. No program will be registered into a system without data to verify it. This is an important early principle because it puts management of software at the same level of importance as management of data within EarthCube.

EarthCube data management infrastructure should establish standard way to register datasets with unique persistent identifiers. These identifiers are used to associate dataset to programs and metadata. These persistent identifiers should be externally resolvable. EarthCube should establish standard techniques for identifying sub-sets and aggregates of datasets.

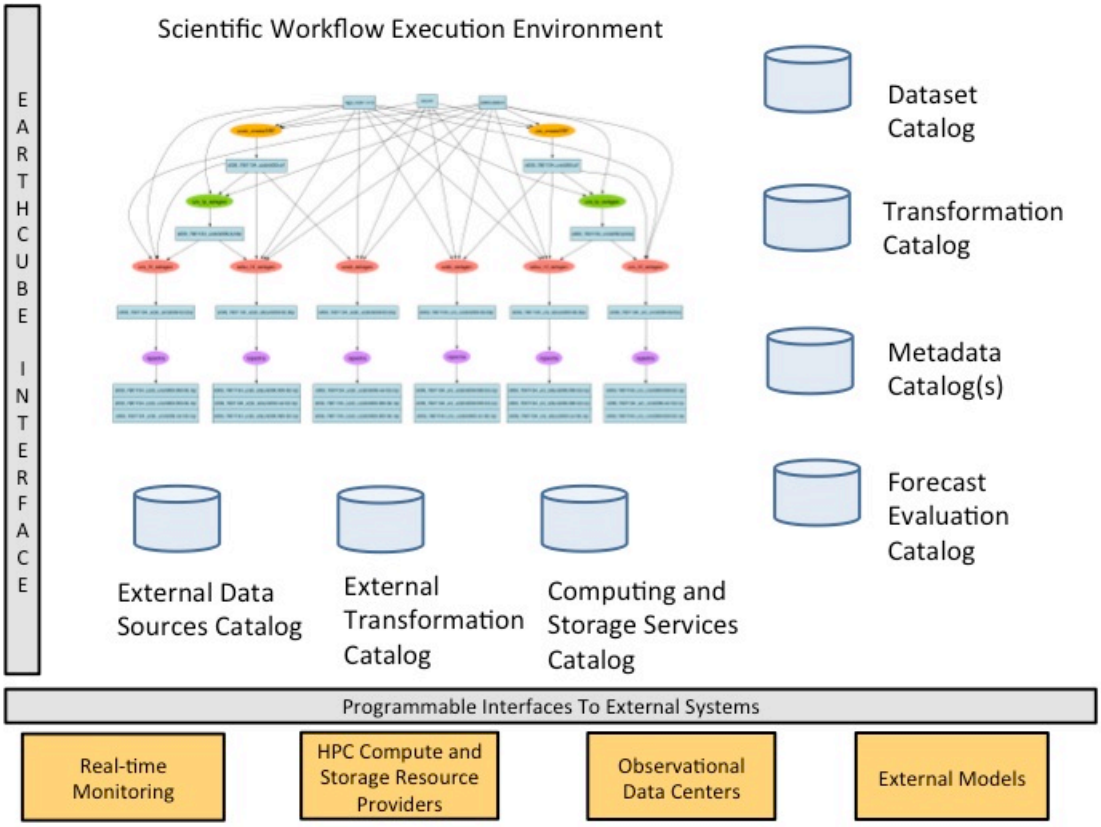

Figure 1: Essential EarthCube cyberinfrastructure includes data management, program management, forecast evaluation, interfaces to other EarthCube systems, interfaces to external systems, and scientific workflow management.

EarthCube software will require a dataset catalog that contains a unique identifier and a mapping to one or more physical files. This will provide the basis for a resolving persistent data ID's and physical files. EarthCube will establish external interfaces to this data catalog so information system can be developed that describe available datasets.

A metadata catalog will associate a unique identifier with metadata about the data. EarthCube metadata management should be extensible, establishing standard ways to support both simple metadata approaches (e.g. attribute-name, attribute-value) and more complex schema-based metadata models. This metadata catalog will contain EarthCube specific attributes that associate datasets with specific transformations and workflows.

EarthCube software management establishes standard methods to register software programs with unique persistent identifiers. Software identifiers are used to associate an executable program (also known as a transformation) with source code, verification data, and metadata.

An EarthCube evaluation system maintains a transformation catalog that defines both forecast processing and forecast evaluation processing. EarthCube provides an external

interface to its transformation catalog. The transformation catalog interface provides information about an EarthCube system's internally, and externally, accessible transformations.

EarthCube forecast processing and forecast evaluations are modeled as scientific workflows. These workflows combine input data, transformations, and output forecasts. Reproducibility of forecast can be achieved through persistent unique identifiers of workflow input data and transformations. Metadata descriptions for forecasts and evaluations produced by an EarthCube system include the workflow involved with identifiers for each data and transformation.

The EarthCube workflow system provides the following basic capabilities. It provides a format for expressing the computational inputs, transformations, and processing order. It provides a workflow engine that imposes data dependencies and processing order. It provides a data catalog, a transformation catalog, metadata management for both, and external interfaces to each of these catalogs. It supports distributed data and computation.

EarthCube provides a data, transformation, and service interface management system. The EarthCube interface management system provides an inventory of available interfaces for accessing data and computing resources. Distributed data and computation is supported by an EarthCube system through interfaces offered by external systems. An EarthCube evaluation system will contain an external interface catalog. The design goal is that external data and external transformation can be referenced within workflows.

EarthCube provides users with tools needed to use existing computational and data management interfaces. Forecasting may require access to very large data stores, and extremely high performance computing. The EarthCube system will provide the interfaces needed for EarthCube forecast and forecast evaluations to use external computational and data storage resources, including web services, NSF HPC resources, and commercial data and computing clouds.

As an evaluation system, EarthCube must manage reference inputs and references solutions. EarthCube will require a specialized knowledge management capability that associates forecasts, forecast evaluation processing, and forecast results.

4) The Design Process: User requirement-driven design methodology, identification of design team members, qualifications of development team, time-line for design demonstration and scale-up, design tools and practices that create robust, sustainable, well-documented and open source infrastructure.

We believe that EarthCube, as an NSF cyberinfrastructure development project, must be organized towards a specific and well-defined purpose to achieve efficient development. Given a specific problem to solve, cyberinfrastructure development can progress rapidly. Without a well-defined objective, wasted development efforts are inevitable. So, EarthCube development should be organized around management of forecast inputs, models and results.

In our work on seismic hazard forecasting, we have identified increasingly rigorous forecast evaluation requirements that depend on the users of the forecast. The progression of computational rigor needed by increasingly broad impact forecast is shown in **Figure 2.**

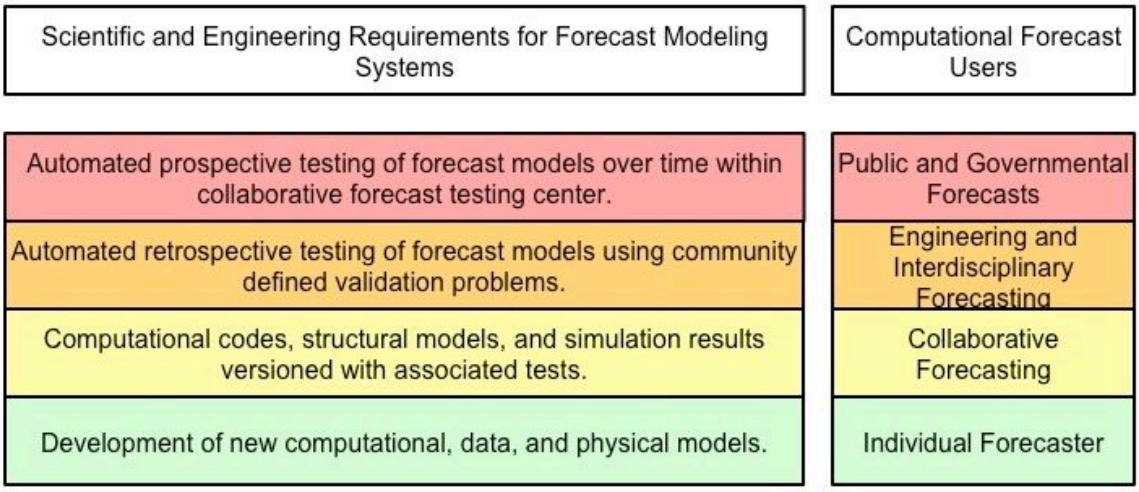

Figure 2: EarthCube cyberinfrastructure should be designed and build from least demanding to most demanding applications. Initial prototypes support individual and collaborative forecasting extended until it can support highly complex and rigorous forecasts.

We propose that the EarthCube design process should be developed from the bottom-up, establishing the simplest and most basic capabilities needed by the broadest number of groups. Once the required tools exist, and are released to the community, increasingly demanding requirements are introduced and new cyberinfrastructure is developed.

An EarthCube prototype should be assembled from a suite of existing tools rather than as monolithic system. EarthCube prototype should be assembled from simple exiting tools. This emphasizes system integration rather than new software development. Monolithic architectures (e.g. everything should be web services) should be avoided until external support for such development is available.

The design process should begin by examining existing forecast evaluation systems. Existing system to be evaluated include PCMCI [1] (climate forecast evaluation), CSEP [2] (short-term earthquake forecast evaluation), and Model Evaluation Tool (MET) [3] (weather model evaluation). These, and other similar systems, should be evaluated for generality, portability, and usability. Capabilities common to all such evaluation systems are factored out as EarthCube system requirements. One, or more of the selected systems is then selected as the basis for a prototype EarthCube. This system is then extended to include the capabilities identified our EarthCube conceptual architecture.

Assuming this approach is adopted, significant efforts will be needed in the following areas.

EarthCube will need to develop forecast specifications and evaluation criteria. For efficiency, we recommend selecting existing forecasts types rather than defining new forecast types.

EarthCube will need to select a small number of representative geoscientific forecasts or forecast elements. EarthCube development is then organized around forecast evaluation, rather than around any specific dataset or data type. Data, and forecast processing, used by the selected forecasts establish the data and transformations that are used to develop the EarthCube prototype. Domains involved in the selected forecast types need to do their own metadata management as representative cases. We support practical metadata techniques based on minimum requirements rather than on all possible future requirements. Each dataset and transformation used in a forecast or evaluation should be associated with sufficient metadata to reproduce results from critical papers.

EarthCube will require collaborative effort to identify prototype interfaces to computational and data resources that might include NSF and cloud resources. The EarthCube design process must establish external interfaces to other EarthCube catalogs. EarthCube design process will develop test cases for data, transformation, and workflow capabilities identified in our conceptual architecture.

The EarthCube design should implement capabilities starting with the lowest-level, most general, capabilities needed to do collaborative research. Then, as the system emerges, and proves useful, it is extended to meet requirements of most demanding forecast types. EarthCube capabilities are expanded to support advanced, broadly used, forecasts only when more general, less demanding, capabilities have been established. The release of general-purpose, research evaluation tools that support individual research and collaborative research will lead to decentralized development with users extending the EarthCube tools for their own purposes.

5) An Operations and Sustainability Model: Operational aspects of a community-wide enterprise need to address such activities as: centralized functions, coordination of services, user services including training, etc. What will it take to sustain an infrastructure that can viable over a long periods of time and who will carry out those functions?

EarthCube must be both scientifically and economically sustainable.

EarthCube scientific sustainability is grounded in its value to scientific research organizations. Unless EarthCube cyberinfrastructure provide value to researchers, it is not sustainable. The EarthCube system capabilities that produce automation and reproducibility are transformative across multiple domains. There exists broad agreement across scientific organizations that well managed data and software is required by modern research. Traceability of results is essential, and coordination of forecast comparison is

essential. EarthCube can become the tools needed to develop and automate forecast evaluation.

The scientific community recognizes the value in introducing scientific rigor into computational science. The EarthCube system is designed to provide unique and persistent identification of both observation and model data, as well as to reduce controversy around forecast results. Because forecast evaluation systems build scientific credibility in forecasts, such systems provide significant value to the scientific community. This is the basis for scientific sustainability.

Our EarthCube economic sustainability model is grounded in the economic value of geoscientific forecasts. Not everyone needs geoscientific forecasts. But in many cases, such forecasts have clear economic value.

As an organization, the EarthCube economic group identifies external, non-geoscientific users of geo-forecasts and solicits support for development, operations, and maintenance of EarthCube. Supporting groups, that might include transportation, insurance, or construction, voluntarily invest in forecast improvements through their own self-interests. Sponsoring organizations invest out of self-interest because they recognize the potential economic value of improved forecasts to their organization.

EarthCube uses support for specific forecast improvements to support development of general EarthCube capabilities. EarthCube capabilities developed to benefit the supporting organizations also provide value to un-related forecasters through open cyberinfrastructure tools.

Forecast improvements provide well-defined metrics for improvement. Quantified improvements to forecasts are used to document value of systems to insure continued funding. As specific forecasts are optimized, further improvements have no further practical value, so rigorous evaluation is no longer performed. The EarthCube forecast lifecycle is completed when no significant improvements to forecasts are expected.

References:

[1] Program for Climate Model Diagnosis and Inter-comparison (PCMDI) http://www-pcmdi.llnl.gov/

[2] Southern California Earthquake Center (SCEC) Collaboratory for the Study of Earthquake Predictability (CSEP): http://www.cseptesting.org

[3] National Center for Atmospheric Research (NCAR) Developmental Testbed Center (DTC): http://www.dtcenter.org/met/users/