EXECUTIVE SUMMARY: EARTHCUBE WORKSHOP RESULTS
EarthCube Modeling Workshop for the Geosciences
April 22-23, 2013 Boulder, Colorado

Organizing Committee & Contributors:
Jennifer Arrigo, CUAHSI
Jed Brown, ANL
Louise Kellogg, CIG UC Davis
Lorraine Hwang, CIG UC Davis
Scott Peckham, CSDMS, University of Colorado, Boulder
David Tarboton, Utah State University
Participants: 55 on site; 7 virtual

Across the geosciences, models of the solid and fluid dynamics and physical processes of the earth and space systems advance our scientific understanding of complex environments and our ability to translate our science into useful societal applications. As EarthCube seeks to develop a data and knowledge management system to transform the geosciences, the input of groups and individuals whom have built-up their own infrastructure and communities around modeling efforts will be critical. While the scientific problems addressed by the broad community of geosciences "modelers" are varied, there are strong commonalities in the computational challenges and requirements of many of these communities that should be exploited to meet these challenges and be a central goal of EarthCube.

This workshop documented the experiences and expertise of well-defined modeling communities within the Geosciences that have, over time, developed their own community, infrastructure and resources. The workshop assessed the needs and readiness of modelers in related geosciences disciplines who do not currently have access to similar resources or community organizations, and provided recommendations that can inform the development of EarthCube.

Science was initially advanced through parallel pillars of theory and observation. With the advent of computation and big data systems additional methods of discovery and knowledge creation involving computation (modeling - the third pillar) and data intensive analysis (the fourth paradigm - Hey et al., 2009) have emerged, and in many cases dominated scientific discovery. While the majority of prior EarthCube domain workshops have focused on knowledge or data management in specific EarthCube/geoscience domains, this workshop examined the role of modeling in contributing to the creation of geoscience knowledge and considered the question as to the modeling infrastructure that should be part of the EarthCube enterprise.

We recommend that cyberinfrastructure that supports modeling should be a key part of the EarthCube cyberinfrastucture as models are an inseparable part of knowledge creation, and model development needs to be curated and formalized much like data management. To give substance to this notion we recommend that EarthCube cultivate the craft of scientific model development and use, or "Model Carpentry" (phrase adapted from www.software-carpentry.org). The workshop identified some specific model development practices that are essential to accelerate the advance and sustainability of models as a pillar for discovery in the earth sciences. These include:

- Community model development. It is imperative that model development practices be structured to facilitate open contribution.
- Abstraction and compartmentalization. Modeling systems are needed to allow questions/programs/models to be framed at a high level, but draw upon bundled CI

services/models/solvers that allow scientist to focus on the science question and let the system take care of the computation and data access. Compartmentalization promotes re-use of components and libraries.

- The social elements of model development. It is critical that there be training and workforce development in support of modeling, career paths for researchers and software developers engaged at disciplinary interfaces, and governance and policies that support collaboration around models.

SCIENCE DRIVERS AND CHALLENGES

1. Important science drivers and challenges: Participants identified several high-priority science questions that will serve as drivers for interdisciplinary modeling efforts in the geosciences during the next 5-15 years.

- How do we integrate and understand multiphysics between highly and weakly coupled systems? e.g. coupled dynamics of fluids, magma, and the solid earth at plate boundaries; co-evolution of hydrologic, geomorphic, critical zone and the deeper subsurface in the face of climate and tectonic drivers.

- How do we integrate and understand the impacts of anthropogenic activity? e.g. feedback between components of the hydrologic cycle, atmosphere, and biosphere and land use and climate change and the role of human activities in these changes and implications for the quality and availability of water for drinking and other uses under increasing demands and scarcity.

- How do we integrate the large degree of spatial and temporal variability in our models? Problems in the geosciences span time scales of $<10^{-6}$ to $>10^{15}$ secs to length scales of $<<10^{-6}$ to $>10^{6}$ m challenging the limits of both methodology and technology. This is unattainable purely by increasing resolution and necessitates the development of multiscaling modeling methods. Methods must account for the translations of variables in time and space, coupling between models, model (non)smoothness and uncertainties (whether numerical or data driven).

- How do we determine model uncertainty and communicate it to both scientists and lay persons? Uncertainty arises from many sources including data generation and assimilation, model limitations, and poorly understood physical processes or processes represented at an aggregate scale using conceptual or empirical parameters. Models are increasingly being used as tools for "engineering" purposes and hence exert influence on policy, resource management, and exploration.

Our workshop did not attempt to develop use cases because of the diversity of problems addressed by models. However, we noted several examples of regions and problems that are closely connected across space and time, and these provide opportunities for synergy across modeling communities. One example (of many) is modeling science in Cascadia (the Pacific Northwest). This region is a locus of intensive study of geology, geophysics, natural hazards (earthquakes, volcanoes, and landslides), landscape evolution, hydrology, climate, and ecosystems, and provides multiple examples of how models link to data integration, modeling on multiple scales, and the dynamics of coupled Earth systems. The Cascadia subduction zone hosts a Long Term Ecological Research Network (LTER) site, is a focus area for GeoPrisms, and has extensive observations from EarthScope's US Array and Plate Boundary Observatory. Modeling is being used to understand problems including the role of fluids in the dynamics

of subduction, and in the evolution of the landscape. These models integrate remote sensing, geochemical, geophysical, and geological data, with the attendant needs and challenges associated with access to data and interdisciplinary communication, many of which have been discussed at other EarthCube end user domain workshops. There are numerous challenges and opportunities for EarthCube that are directly associated with data acquisition, assimilation, and modeling in such cross-cutting regions or topics of study.

2. Current challenges to high-impact, interdisciplinary science: Several themes emerged as consistent challenges faced within/across the involved discipline(s) (list 3 to 6).

1. Language and interaction. Individual disciplines each have their own community vocabularies, language and expertise, posing challenges for those working within and across disciplines on interdisciplinary science. Some communities have developed either formal or informal standard names for variables or processes that are immediately understood by others in their discipline; other communities may have several terms or ways of describing concepts. Within disciplines, these concepts and terms are well understood, because implicit in the terms are an understanding of the science and context. However, to do interdisciplinary science, information, data and models must be shared and understood ***across*** disciplines, and we cannot depend on this implicit understanding. This is both a ***technical*** challenge (in terms of metadata for both data and models, interoperability and assessing fitness for use across disciplines) and a ***social*** challenge (in terms of scientists being able to share knowledge and work effectively across disciplines, and for scientists, engineers and mathematicians to work on common problems).

2. Challenges surrounding open access and sharing of codes, models and software. Participants largely felt that open access and sharing is important for interdisciplinary science and collaboration, but there are many unresolved issues and questions even within modeling disciplines, including (but not limited to):

- credit and recognition for contributions (data, models and software) within the current scholarly reward structure.
- questions of ownership and provenance of models, code, techniques, algorithms, and software.
- how to adequately describe a model and its limitations so that others can assess and use it. This includes worries about model misuse (intentional or unintentional) by others. We note that some end user domain workshops expressed a wish for easy-to-use modeling codes, while the modeling community, who actually develops models, is more cautious, and wants to see appropriate training, documentation, and awareness of the strengths and limitations of models.
- the burden of supporting a code once it has been released to the community. Some communities (e.g. atmospheric sciences, geodynamics, surface processes) have extensive support for community models, which can include community code repositories, dedicated staff and resources for managing and maintaining the code. Interestingly, the cyberinfrastructure used by these communities have both similarities and differences that reflect the needs of the scientific domains. Moreover, individual researchers that have developed models and codes that they are willing to share often do not have the time, resources or desire to provide such “operational” support. This inhibits re-use of code and sharing of knowledge.

3. Diverse types and approaches to modeling for different purposes. Models are an abstraction of reality to focus on a specific problem of interest; each model is developed with a specific purpose. The purpose drives the way that the physical environment is described, and may include simulation of a physical system, exploration of the physics of a problem through exploration of the effect of the controlling parameters, or investigation of the stochastic behavior of a system to understand possible behaviors or

states of that system. Depending on the purpose, the process may be represented as a suite of partial differential equations (PDEs) to be approximated numerically, or as more aggregate or lumped objects that represent discrete components of a system. A simple example of this distinction is Geographic Information Systems (GIS), where information may be represented using discrete shapes (point, line, polygon) in geographic space, or on grids that represent information at the scale of the grid. In developing new algorithms and models, the researcher must determine whether an object-based or PDE-based approach is optimal. A challenge is to develop computational frameworks that integrate reductionist and object approaches or deterministic versus stochastic approaches. The deterministic versus stochastic approaches also need consideration in evaluating results. It remains a challenge to determine whether a model should match observations as closely as possible, or only in a statistical or regime sense; the answer tends to vary from problem to problem and even from researcher to researcher.

TECHNICAL INFORMATION, ISSUES, and CHALLENGES

1. Desired tools, databases, etc. needed for pursuing key science questions with brief elaboration:

Algorithm development: In parallel to the advances in computational hardware power, advances in algorithms, software, and compilers enable better, more effective use of advanced computing. Optimal algorithms become more critical as we solve larger problems on larger computers. Continued advances require support for developing portable mathematical and numerical methodologies across fields of geoscience. New methods require research by applied mathematicians, computational scientists, and statisticians (among others) that is motivated by geoscience problems.

In addition to research advances, implementation of new algorithms requires skilled, experienced software engineers to develop and support community codes and assist geoscience researchers with code development. However, it can be difficult to recruit and support software engineers in the domain sciences; it is essential that attractive career paths and sustained support be available to talented software developers. The challenges and barriers to new algorithms include sustained support for both ends of this spectrum (research, and code development and hardening.)

Visualization: Scientific visualization is an essential element of the scientific work for modeling. Models can generate very large, complex, and high dimensional data; scientific visualization is a fundamental tool for analysis of these data, extraction of features, data assimilation, verification and validation of numerical methods, and extracting insight. Scientific visualization is used as a preprocessing aid to assemble inputs and discretizations for models. Finally, scientific visualization is used to communicate results and discoveries to the research community and beyond, to policy makers, educators, and the general public. The technical challenges and issues include availability of adequate methods for visualizing complex and diverse data types, integration of visualization at all appropriate steps in the workflow, visualization of very large datasets, and adaptation of new technologies.

Models: Infrastructure is needed to support model reproducibility, reusability and transparency. Community models require sustained development and support and community tools for working with them, such as workflows and software for managing the enormous amount of scientific and computational choices that go into models. Community standards for testing, computing and portability of model codes would greatly enhance the impact of these models. These standards would aid in the creation of more flexible and easier to use community models, and would enable more effective science in a research environment that has a rapid pace of scientific and technological development, limited resources for developing and sustaining meaningful collaborations, and an existing and enormous diversity in model

structures, programming languages, computational platforms and data requirements. Such models should seamlessly access data resources and parameters.

Advanced computing: Modeling typically requires access to advanced computing resources, including (but not limited to) large-scale high performance computers such as are available from the Yellowstone-NCAR-Wyoming facility, NSF's XSEDE facility, and leadership class DOE computers. Advanced computing may also include mesoscale parallel computing, from small clusters operated by individual PIs to mid-sized clusters; these can be difficult for PIs to obtain and operate. Modeling science requires effective access to and assistance using such computing facilities, in order to make best use of the investments in computing hardware. New technologies (such as GPUs) are emerging, requiring re-development of models to take advantage of increases in performance.

Model and Data uncertainty: As multi-disciplinary efforts emerge to model multi-scale and long-term processes, researchers are challenged to identify systematic and rigorous ways to rapidly assimilate new data and to characterize the statistical structure of observational data. It is important to pay attention to systematic, random, and model error as well as possible sources of unknown errors. For even well-understood systems, predictive modeling with quantified uncertainty and model-based experimental design places new demands on characterization of uncertainty in both observational data and models. For less well-understood systems, different approaches must be explored. These different sources of error and uncertainty are not currently well-communicated, and to the extent that such communication takes place, it is usually only within a community or scientific domain, and not beyond. Communication of uncertainty is especially important for those who must try to craft policy from science. Since uncertainty quantification is an active area of research containing many open theoretical, methodological, and algorithmic questions, one challenge is ensuring that methodology and cyberinfrastructure be made extensible in order to support future innovations.

COMMUNITY NEXT STEPS

*What the community needs to do next to move forward and how it can use EarthCube to achieve those goals***:**

Recommendations:

1. Support and resources for interdisciplinary research partnerships for geoscientists with applied mathematicians, statisticians, computational scientists, computer scientists, and the like. These collaborations are essential to advance methodologies used for modeling, and will provide a foundation for the next generation of computational methods for the geosciences. Such collaborations are also necessary to develop statistical models of uncertainty in observational data, and methods for propagating uncertainty through models; these models and methods are likely to emerge as a core component of observational data provenance. An example of one (past) mechanism for doing this was NSF's solicitation for Collaborations in Mathematical Geosciences (CMG), now closed. This program resulted in successful, productive collaborations between geoscientists and mathematical scientists, with research advances in both disciplines that have been incorporated into geoscience modeling.

2. Mechanisms to support ongoing dialogue and intensive interdisciplinary collaboration. Interdisciplinary research requires ongoing communication among groups (large and small), through workshops, forums, remote collaboration tools, and other tools. EarthCube should facilitate development of communication and collaboration tools that are seamlessly integrated with the data and modeling infrastructure of EarthCube, to provide effective "workspaces" for groups in addition to communication.

3. Advanced computing: Modeling geosystems at the highest resolution requires effective access to mid-scale parallel computing, leadership class high performance computing, and associated advanced computing tools. HPC resources are available through investments by NSF and other federal agencies; however, for individual researchers, an effective pathway from desktop computing to HPC remains challenging. The pathway to using advanced computing resources requires an investment in computational scientists who can work closely with domain scientists to achieve their goals; development of high-quality codes using best available methods, as well as tools for managing and analysing the data that emerges from models, including scientific visualization, and access to mid-scale computing. Projects developing software should be encouraged to adopt workflow and design practices that will foster community involvement and upstreaming of contributions.

4. Training and education: Scientific advance using models depends on a cyber-enabled workforce of researchers who understand both the geoscience domain and the mathematical and computational foundations used for modeling. It is therefore critical that there be training and workforce development in support of modeling.

5. Social and cultural changes: A cultural change is needed to enable scientists to facilitate open access to data and ensure that scientists receive credit for their work. Although we do not have a solution to this problem, we see a timely opportunity for NSF to investigate possible solutions, in conjunction with the move to open access of data, model results, and codes. The EarthCube community can make an important contribution to this dialog. EarthCube also should support the development of technology and approaches that address product (e.g. model, code, and software) citation, description, provenance, and related issues that could form the basis of infrastructure that would be needed in conjunction with the cultural and social changes. As noted above, methodology and cyberinfrastructure must be extensible in order to support future innovations.

Reference

Hey, T., S. Tansley and K. Tolle, (2009), The Fourth Paradigm, Data-Intensive Scientific Discovery, Microsoft Research, Redmond, Washington, 283 p, http://research.microsoft.com/en-us/collaboration/fourthparadigm/.