EarthCube End User Workshop Executive Summaries

Compiled by EarthCube Office Staff
Last updated 12 December 2014

Note: When referring to page numbers in this document, please reference "Compilation page" numbers.

Contents

Introduction

In order to reach out to potential end-users of the National Science Foundation's (NSF) EarthCube initiative, NSF funded a series of two dozen EarthCube domain end-user workshops throughout mid-2012 to late 2013, targeting a broad spectrum of Earth, atmosphere, ocean, and related scientists, including senior and early career scientists. The purpose of these workshops was to allow geoscience communities to articulate and document their cyberinfrastructure needs and what they would like to do in the future, in terms of accessing data and information within and outside their disciplines.

A specific goal of these workshops was to gather requirements on EarthCube science-drivers, data utilities, user-interfaces, modeling software, tools, and other needs so that EarthCube can be designed to help geoscientists more easily do the science they want and need to do. More specifically, science that helps address the NSF's GEO Vision, 2009: fostering a sustainable future through a better understanding of our complex and changing planet.

This document is a collection of all end-user workshop executive summaries as collected and compiled by the EarthCube Office.

EXECUTIVE SUMMARY:
EARTHCUBE END-USER DOMAIN WORKSHOP REPORT

Submitted: 13 November 2013

Editors: Planning Committee Members

Workshop Title:
Articulating Cyberinfrastructure Needs
of the Ocean Ecosystem Dynamics Community

Woods Hole Oceanographic Institution
Woods Hole, MA
7-8 October 2013

PLANNING COMMITTEE

Danie Kinkade, (Chair) Woods Hole Oceanographic Institution (WHOI)
Cynthia Chandler (WHOI)
David Glover (WHOI)
Robert Groman (WHOI)
David Kline (Scripps Institution of Oceanography, University of California, San Diego)
Jasmine Nahorniak (Oregon State University)
Todd O'Brien (NOAA, National Marine Fisheries Service)
Mary Jane Perry (School of Marine Sciences, University of Maine)
James Pierson (University of Maryland Center for Environmental Science)
Peter Wiebe (WHOI)

INTRODUCTION

An EarthCube Water Column Domain End-User Workshop hosted by the Biological and Chemical Oceanographic Data Management Office (BCO-DMO) was held October 7-8, 2013, at Woods Hole Oceanographic Institution. This executive summary synthesizes the workshop discussions, while further details such as use case descriptions and results of plenary and breakout group discussions are provided in the full workshop report.

The goal of the workshop was to articulate cyberinfrastructure needs of the ocean ecosystem dynamics community with particular focus on the challenges presented by multi-disciplinary marine ecosystem research. The ocean ecosystem dynamics domain encompasses a broad array of disciplines that often requires investigations in four dimensions, and seeks to increase understanding of the interplay between biological, chemical, and physical processes in the ocean. It is fundamentally an interdisciplinary domain by nature, producing highly diverse data types that pose unique challenges for management, integration, and analysis. The ability to discover, access, and synthesize high quality data from various disciplines is crucial to ocean ecosystem sciences.

The workshop brought together 50 participants (43 on-site and 7 remote) to explore and document the community's cyberinfrastructure needs from the domain users' viewpoint. The participants included 22 established, 16 mid-career, and 12 early-career or postdoctoral researchers. Individuals self-identified into one or more disciplines: oceanographers (68% of total), data and information managers (42%), cyberinfrastructure researchers (22%), education and social science specialists (14%), and modelers (8%). It should be noted that the timing of the lack in Federal appropriations and subsequent government shutdown prevented several registrants from participating in this workshop.

SCIENCE ISSUES AND CHALLENGES

1. Important science drivers and challenges: Participants identified several high-priority science questions that will drive interdisciplinary research efforts during the next 5-15 years.

- How will ocean acidification, warming, and hypoxia affect marine ecosystems? How do these phenomena affect the organisms? What species will be most impacted? What will be the impacts on ecosystem structure and dynamics?
- As ocean circulation changes in response to warming and changing salinity, how will marine organisms respond to resultant changes in their environment? What will be the impacts on productivity, species distributions, and carbon flux in the ocean? How will such ecosystem changes affect human populations?
- How will species distributions, life histories, and species interactions in polar oceans be impacted by the changes in ice cover? How will these polar changes impact other regions including changes in weather patterns? What will be the bottom up effects on marine coastal and open ocean ecosystems?
- How do different physical, chemical, and biological processes come together to create more complex emergent properties in ocean ecosystems? What are the feedback loops among these processes and how do they give rise to biological, chemical, and genetic diversity in the oceans?
- How are anthropogenic impacts other than climate change, such as eutrophication, overfishing, and coastal development influencing marine ecosystems? How does this influence our approach to ecosystem based management and conservation?

2. Current challenges to high-impact, interdisciplinary science:

A. Participants recognized three different categories of barriers preventing them from conducting transformative science: cultural, institutional, and technical.

Cultural barriers highlighted the reluctance to share data, a lack of confidence that data will be cited, different sharing expectations across domains, and the inability to access dark data (defined below in section B). Also mentioned were differences between domain vocabularies and an inability to translate science to the public.

Institutional barriers include the existence of organizational stove pipes that hinder cross-discipline collaboration, and a lack of the following: inter/intra-agency collaboration on data management, incentive and support for international and interdisciplinary collaborations and proposal reviews, recognition by tenure committees when submitting data for sharing, support for existing routine monitoring and observations, and insufficient funding for cutting edge science.

Technical barriers focused on aspects of data, such as lack of a single source or common method for discovery and access, and a lack of quality control assessments. Other data related barriers included difficulty with large data sets, combining heterogeneous data from different types of data streams, and difficulty dealing with evolving data types (such as 'Omics' data). Additional technical barriers were a lack of analysis code/tools for automation and visualization, and a lack of effective technological solutions to share data, etc.

B. Additional, more detailed challenges faced by participants:

Databases Difficult to keep on top of all current cyberinfrastructure efforts (DataONE, EarthCube, Data Conservancy); lack of interoperability and diversity of data structures among data centers; lack of a common database capable of storing, searching, and serving terabytes of images and videos; submission of the same data to multiple repositories; difficulty knowing where to submit data.

Access Lacking clearinghouses for (a) software/tools/code (b) databases (c) models, and (d) vocabularies/ontologies; difficult to find, access, and work with data outside of your area of expertise; difficult to access data from international databases; difficult if not impossible to access data from publications.

Dark data No access to dark data (data on local computers or in file cabinets that are not in any public repository and not discoverable); challenges include lack of metadata, changing personnel, and lack of time and funding.

Data types Some data are difficult to put into repositories (due to size, structure, format).

Metadata Lack of rich, standardized metadata.

Quality No standardized way to assess data quality

Tools Lacking tools for automation, integration of a variety of data, and visualization in 4-D; a challenge to combine different data types and to deal with new and evolving data types; difficult to analyze heterogeneous data with gaps; a lack of systems for documenting provenance.

Hardware Increasing need for data storage, processing capabilities, and bandwidth.

Education Lack of data and computational literacy; need for data literacy courses; lack of online training tools for discovering, accessing, and using data.

Effort Substantial time and cost in the effort of making data available. Reviewers are often unsympathetic to the high cost of data management activities.

TECHNICAL INFORMATION/ISSUES/CHALLENGES

Identified below are several critically needed tools, repositories, and infrastructure needed for pursuing key science questions:

1. Tools:

- Tools to enhance data discovery and searches.
- Visualization tools for interactive data analysis.
- Tools to track and control data versions.
- Tools to foster data quality assurance and quality control.
- A community-level interface and facility to share tools and programming code.
- Technology to assimilate metadata produced by smart sensors.
- A tool to translate from different format types to a standard format.
- Automatic incorporation of new data into databases/repositories
- Automatic retrieval of data for use in applications (e.g., forecasting)

2. Repositories and Databases:

- New data repositories (or expansion of existing facilities) are needed for emerging data streams that currently are not supported by a repository (e.g., metabolomics, citizen science data)
- A system for handling massive amounts of data including images and video.
- Production of a wider range of more sophisticated data products and derived calculations.
- Ensuring repositories function as, or work with archive facilities for long-term preservation of data.

3. Global Infrastructure:

- Guidance on where to submit data including restrictions and guidance for repository use.
- A centralized forum providing information about models, scripts, software, and documentation.
- Undergraduate and graduate-level curricula for training the next-generation of scientists to be able to find, submit, and work with data.
- Educate programmers to understand science.
- Standard interdisciplinary metadata format.
- Cross-domain ontologies of measurable phenomena and instrument types.
- Further development of crowd-sourcing funding and technology

COMMUNITY NEXT STEPS

Below is a list of tangible items and actions by EarthCube that would facilitate the community achieving its transformative science goals:

Short-term Next Steps (1-3 years):

- Catalog of different data repositories and tools.
- Catalog of existing and dark tools (for discovering, analyzing, and visualizing data).
- A tool that captures the output from multiple Earth System models for a geographical position for comparison to field-collected data.
- Better search tools for data discovery (such as faceted searches).
- Help desk to provide investigators with information on repositories where data should be submitted.
- Provide incentives to preserve all data and make accessible.
- Data and computational literacy curriculum including tutorials for undergraduate and graduates.
- Synthesis of outcomes from the domain-specific EarthCube workshops.

Long-term Next Steps (>3 years):

- Centralized access to earth system and ocean model output data with Google style searchability.
- Making databases and repositories interoperable.
- Plan to identify and liberate dark data (resources, funding, and expertise).
- Provide funds to enable Use Cases put forward in the workshop to be implemented.

WORKSHOP SUMMARY:
EarthCube Domain End-user Workshop: Bringing Geochronology into the EarthCube Framework

Conveners: Brad Singer, Shanan Peters
Co-organizers: Andrea Dutton, Rebecca Flowers, George Gehrels, Brent Goehring, Tom Guilderson, Anthony Koppers, Noah McLean, Stephen Meyers, Susan Zimmerman

Seventy on-site as well as at least eight off-site participants, representing a range of geochronology sub-disciplines and end-users of geochronology data, gathered in Madison, Wisconsin on October 1-3, 2013. This is the first meeting in the U.S. that has brought together such a large spectrum of geochronologists, whose expertise spans from near-modern to early Earth timescales. We discussed the five NSF-prompted workshop goals below from within- as well as across-discipline perspectives. We recognized that the diverse geochronology communities share many common obstacles and needs, and that each sub-discipline is at various stages of envisioning and developing domain-specific organizational tools and cyberinfrastructure that would feed into a wider EarthCube framework. There is also recognition that investment in cyberinfrastructure has the potential to improve access to high-quality dates, models, age calculation tools, and recalculation tools that will benefit geochronologists as well as the many end-users of geochronology data.

In addition to the summaries provided below regarding each of the desired workshop outcomes, there are several appendices, including: a) a list of workshop participants and affiliations; b) an inventory of current community cyber-infrastructure resources; c) specific data and cyberinfrastructure needs for the geochronology community.

Grand Challenge:

Develop a fully integrated four-dimensional digital earth, of which geochronology provides the crucial fourth dimension, to fully understand dynamic earth system evolution.

Outcome 1: Science Drivers

The primary scientific driver identified during the workshop if EarthCube were successful would be to **understand** and **test** hypotheses about the underlying controls on, and the relationships between, major earth systems. Achievement of this goal will entail establishing:

- a robust, unified chronological **framework** for all earth history;
- **correlation** of earth system records across a range of nested spatio-temporal scales;
- **causality** between forcing, responses and feedbacks, including leads and lags;
- **rates** of change of fundamental earth system processes.

The above provides a general framework for the goals of the geochronologic community within EarthCube. Specific examples of scientific opportunities and challenges facing the geochronologic community over the next 15 years that will lead to resolution of the dynamic interactions among Earth systems include, but are not limited to:

- the construction of a digital absolute geologic time scale used to resolve the times and drivers for biologic extinction as well as the rates of biologic recovery and evolution;
- the pace, magnitude and drivers of climate change through earth history (e.g., the carbon cycle, oxygen, sea level, ocean chemistry);
- addition of a 4th dimension to the construction and evolution of the North American continent, providing knowledge products to be directly integrated with EarthScope data;
- resolving interactions between rates, patterns, and magnitudes of erosion, landscape evolution, and sediment deposition, with climate change and tectonics in deep and more recent time.

Outcome 2: Data & Cyberinfrastructure obstacles
To specifically address the overarching vision of EarthCube, the geochronologic community must overcome several major and minor obstacles that require financial and cyber-infrastructure support. We also identified social/structural obstacles to success. These include:
Data & Cyberinfrastructure obstacles

- Geochronological data are currently difficult to access, of variable quality, and challenging to compare between labs/methods and with other information;
- Limited standardization of data acquisition, archiving and delivery protocols across the geochronologic community;
- The need to archive legacy data and develop mechanisms for managing the current data explosion, so that new and existing data can be leveraged;
- Domain-specific data architectures are vastly incomplete or absent altogether and require the development and maintenance of software designed for the reduction and archiving of geochronologic data, designed to remain flexible for unanticipated data additions;
- There is a lack of transformative technology for integrating earth system knowledge -- data at present are locked in domain-specific architectures;
- It is difficult to recognize gaps and data deserts in existing datasets, and disconnects between disparate datasets;
- To create geochronology data that is amalgamated into databases directly comparable requires financial support of EARTHTIME-like initiatives for various geochronometers to establish community-wide protocols, and evaluate and improve inter-laboratory comparison;
- It is challenging to develop ***continuums*** across human and geologic timescales;
- Educational content for EarthCube users that includes support for preparing the next generation of geochronologists to benefit from EarthCube's "big data world";
- A need for visualization tools that make EarthCube accessible to the non-specialist audience (e.g., non-specialist scientists, K-12 teachers, policy makers).

Social/structural obstacles

- The need for clear mechanisms for defining data ownership and credit pre- and post-publication, which must be addressed for the community to "buy-in" to EarthCube;
- A need for benefits to individual geochronology labs and research groups to motivate their contributions to an EarthCube database;
- Improved community and institutional appreciation of the importance of and opportunities presented by cyber-infrastructure is required to promote widespread adoption of

cyber-based technique- and domain-specific tools.

Outcome 3: Existing community data and cyber resources (see Appendix 2 below)
Existing community data and cyber resources immediately relevant to the geochronological community, as well as datasets that geochronologists leverage for their larger research endeavors are summarized in an appendix. The list indicates that while some geochronologic communities are relatively well-organized in a cyber sense, others are not, but a common thread shared by all are the usage of data- and cyber-resources across the broad spectrum of Earth, Atmospheric, and Ocean sciences.

Outcome 4: Data and cyber-capabilities required
As EarthCube evolves, there needs to be a development and maturation of technique-specific to science application-specific data systems. Notably, it was identified that there is a:

- High priority for Earthcube to help communities develop their own domain-specific data-handling systems, from top-down system design to bottom-up assimilation of existing databases;
- Need for continuity in funding for cyber- and geochronological-infrastructure, including personnel;
- Need to develop expertise, communication and collaboration across a spectrum from cyber-savvy geoscientists to Earth science-dedicated computer scientists -- Marry computer scientists into the Earth science community;
- Need for development and maturation of geochronologic technique-specific to science application-specific data systems, including:
    - Tools that feed data into cyber-infrastructure, including novel systems for automating and easing the input of data;
    - Tools that extract, analyze, visualize, and integrate knowledge from Data Systems (EarthChem/Geochron, UNAVCO, EarthScope, NOAA, Neotoma) to be used within EarthCube;
    - Metadata capture adequate for automated data revisions (e.g. decay constants, reference materials);
    - Snapshots of the database (i.e. legacy) – record of previous versions.

In addition to data- and cyber-capabilities required above, there is also a general consensus that linkage to the publishing domain is important to ensure proper attribution and citation of data and data products (e.g., DOI). The establishment of working groups to discuss common protocols and community standardization, both during the development of technique-specific cyber-infrastructure and afterwards in the usage phase.

Additional required data- and cyber-capabilities identified are summarized in **Appendix 3** below.

Outcome 5: Opportunities achievable with EarthCube development and support
Three grand opportunities provided by the development and maturation of geochronological related data- and cyber-infrastructure within EarthCube are summarized below. Each is unique in its

approach and each addresses problems of such large scale that they are largely intractable in the absence of a unified approach to data integration, such as that envisioned by EarthCube. In 1988, Claude Allegre alluded to the challenges of reconstructing the complex history of the continental crust with a "statistical approach" and that we should "abandon any hope… of a cartographic synthesis". We however, are more optimistic that aspects of these complex problems could be addressed in the near future if diverse emerging and existing datasets are fully integrated in EarthCube.

Potential EarthCube Deliverable #1

A **fully digital geological time scale** is a geoinformatics knowledge product that merges all stratigraphic and chronologic records of the Earth's sedimentary carapace, from the section to basin to global scale, and thus accurately expresses all of the embedded proxies of Earth systems evolution (paleoclimate, paleobiology, critical zone interaction, landscape evolution, basin dynamics, plate tectonics) in a quantitative ordinal framework. The digital geological time scale can only emerge through the federation of multiple domain science data systems, and will provide conclusive tests of a myriad of hypotheses centered around correlation, causality and rates of Earth systems phenomena and processes.

Example outcome: The orbital versus tectonomagmatic control on extreme and/or rapid climate events remains an outstanding question in Earth systems analysis. Geochronology is uniquely suited to testing associated hypotheses that rely on **correlation** of the proxy records that contain the signal of climate change; that predict **causality** between forcing, response and feedback; and that distinguish between alternative hypotheses that predict contrasting **rates** of forcing and response.

Potential EarthCube Deliverable #2

A quantitative model of the **4-dimensional evolution of the Earth's lithosphere** requires the integration of paleogeographic reconstructions, proxy records of paleoelevation and relief, landscape evolution models, and thermal and geochemical constraints on crustal volume and structure. Geochronology and thermochronology play the major role in correlating and calibrating these proxy reconstructions and models. Existing efforts toward this goal have been limited to the basin or orogen scale; EarthCube cyberinfrastructure would enable the means to generate the first continental and global 4D Earth lithosphere models through time.

Example outcome: Existing plate tectonic reconstructions are limited by the lack of extant oceanic plates older than ca. 200 million years, or less than 5% of Earth's evolution. However, the continents preserve signals of plate interaction in orogens, basins and magmatism that have been used to reconstruct more ancient plate configurations in a piecemeal way (e.g. the "supercontinental focus") since the birth of plate tectonic theory. Yet these signals are nearly impossible to consolidate into a global reconstruction using existing methods of compilation and synthesis. **Global plate tectonic paleoreconstructions**—and the very existence of plate tectonics across Earth history—could be an emergent phenomena out of an integrated 4-D digital Earth model tracking ancient plate interactions signaled by synchronous orogenic and magmatic phenomena recorded in now separated continental landmasses.

Example outcome: Elevation and relief impose first-order constraints on atmospheric circulation and modern climate dynamics. Similarly the reconstruction of paleotopography is required to provide boundary conditions for both regional climate and global circulation models seeking to reproduce "alternative Earth" climate scenarios present in deep time, for example the Paleocene-Eocene Thermal Maximum or Neoproterozoic Snowball Earth states. **Paleotopography** is a challenging reconstruction that integrates disciplines as diverse as tectonophysics, structural geology, basin analysis, paleoecology, and stable isotope geochemistry of paleosols and fossils. All of these domain sciences are linked to together by geochronology and thermochronology, whether through direct dating or correlation via the geologic time scale. A 4-D model of Earth's lithosphere could provide quantitative and reproducible paleo-topographic reconstructions of continental landmasses through time that can be used as input for modeling of climate, paleoecological response, sediment dispersal, paleohydrology, and terrestrial geochemical fluxes.

Potential EarthCube Deliverable #3

The recognized need for synthesis of paleoclimate data is a particularly pertinent issue with respect to **sea-level change**. Few efforts have been undertaken to integrate paleo sea-level data in a systematic and rigorous fashion. Indeed, the assimilation of data across multiple timescales with differing chronometers is without precedent. This hampers the interpretation of paleo sea levels on a regional scale and limits the possibilities to tune and refine models that predict sea-level change and its spatial variability and to produce a global sea-level curve.

The development of a global sea-level curve over the last full glacial cycle with the most up-to-date geochronological control requires the integration, standardization, and recalculation of thousands of different individual sea-level constraints spanning multiple chronometers (U-series, C-14). Interpretation of this sea level curve is largely meaningless in the absence of simultaneous correlation with other proxy records (e.g., ice cores, marine stable isotope records), timing of terrestrial ice sheet/glacier change (e.g., via radiocarbon or surface exposure dating), and astrochronology. The ability to establish a robust geochronological framework will allow for accurate establishment of correlation between different locations, as well as different records, determine causality of sea-level changes, including leads and lags, and determine rates of sea-level change.

Example outcome: We envision the development of a user interface to extract ages of sea-level markers (such as U-series ages and C-14) and update and normalize these ages into modern calibrations (e.g., CALIB13, updated decay constants). These updated data would be fed into a domain specific database (SeaBase) for sea-level markers and subsequently fed into existing glacial isostatic adjustment (GIA) ice models to refine their temporal framework. Combining and iterating the GIA model with the newly developed dataset would enable construction of a best-estimate global (eustatic) sea-level curve over the last glacial cycle based on absolute dates of direct markers of sea level.

While the above are grand challenges faced by the community and possibly viable with an EarthCube, **immediate next steps** have also been identified and are summarized below.

- Highest Priority: Development and maturation of domain-specific cyber-infrastructure to broaden and democratize participation.
- Dissemination activities within each sub-discipline to foster national/international community buy-in and participation
- Town-hall at GSA Denver (October 2013)
- Communication with NROES committee on Geochronology
- Identify and develop RCN opportunities
    - Within the geochronology domain
    - Across domains: integrate/link with existing EarthCube domains

Appendix 1. Workshop Participants

Name		Affiliation
Sarah	Aciego	University of Michigan
William	Amidon	Middlebury College
Nathan	Andersen	University of Wisconsin-Madison
Jason	Ash	IEDA Group / University of Kansas
Yemane	Asmerom	University of New Mexico
Greg	Balco	Berkeley Geochronology Center
Andrew	Barth	Indiana University
Erin	Birsic	University of Wisconsin - Madison
Terrence	Blackburn	Carnegie Institution for Science
Kimberly	Blisniuk	University of California, Berkeley
James	Bowring	College of Charleston
Samuel	Bowring	MIT
Andy	Calvert	US Geological Survey
James	Channell	University of Florida
Drew	Coleman	University of North Carolina
John	Cottle	University of California, Santa Barbara
Andrew	Cyr	U.S. Geological Survey
John	Czaplewski	University of Wisconsin - Madison
Andrea	Dutton	University of Florida
Alison	Duvall	University of Washington

Lang	Farmer	University of Colorado, Boulder
Rebecca	Flowers	University of Colorado Boulder
Julie	Fosdick	Indiana University
George	Gehrels	University of Arizona
Brent	Goehring	Purdue University
Simon	Goring	University of Wisconsin - Madison
Eric	Grimm	Illinois State Museum
Thomas	Guilderson	LLCenter for Accelerator Mass Spectrometry
Sidney	Hemming	Columbia University
Timothy	Herbert	Dept. of Geological Sciences, Brown University
Jon	Husson	Princeton University
Brian	Jicha	University of Wisconsin-Madison
Shari	Kelley	New Mexico Bureau of Geology and Mineral Resources
Anthony	Koppers	CEOAS, Oregon State University
Todd	LaMaskin	University of North Carolina Wilmington
Todd	LaMaskin	UNC-Wilmington
Thomas	Lapen	University of Houston
Alberto	Malinverno	Lamont-Doherty Earth Observatory of Columbia University
Shaun	Marcott	Oregon State University
Michael	McClennen	University of Wisconsin-Madison
David	McGee	MIT
Noah	McLean	University of Kansas
Steve	Meyers	University of Wisconsin - Madison
Brent	Miller	Texas A&M University
Brent	Miller	Texas A&M University
Kyoungwon Kyle	Min	University of Florida
Andreas	Moeller	University of Kansas - Department of Geology

Leah	Morgan	Scottish Universities Environmental Research Centre
Roland	Mundil	Berkeley Geochronology Center
Bette	Otto-Bliesner	National Center for Atmospheric Research
Lisa	Park Boush	University of Akron
Genevieve	Pearthree	Arizona Geological Survey
Shanan	Peters	University of Wisconsin - Madison
Troy	Rasbury	Stony Brook University
Ken	Rubin	Univ. of Hawaii, Dept. of Geology and Geophysics
Brad	Sageman	Northwestern University
Allen	Schaen	University of Wisconsin - Madison
Mark	Schmitz	Boise State University
Blair	Schoene	Princeton University
Brad	Singer	University of Wisconsin - Madison
Keith	Sircombe	Geoscience Australia
Keith	Sircombe	CSIRO Australia
Stuart	Thomson	University of Arizona
Basil	Tikoff	University of Wisconsin - Madison
Laura	Webb	University of Vermont, Department of Geology
Jack	Williams	University of Wisconsin-Madison
Susan	Zimmerman	Center for AMS, Lawrence Livermore National Lab

<u>Virtual Participants</u>

David	Anderson	NOAA
Daniel	Condon	British Geological Survey, NIGL-NERC
Kip	Hodges	Arizona State University
Matt	Horstwood	British Geological Survey NIGL-NERC

Appendix 2: Outcome 3 - Existing community data and cyber resources

Multi-disciplinary

Databases

National Map (USGS) – GIS data and map layer files
National Mapping Center (USGS)
USGS online reports
State and country map and fault repositories
National Geochronology Database (USGS)
Supplementary data to published papers (GSA, ESA, PNAS, Science, Nature, …)
ProQuest (theses)
Earthchem/IEDA
NAVDAT
GERM – some standards data (actively updated).
USAP (US Antarctic Program) –
GNS new zealand dem and rock chem database.
NASA databases
GeoRoc – General geochemical/isotope DB (Global)
TephraBase/USGS Rock/
PetDB petrological data

Software

GeoDeepDive

U-Pb, Ar/Ar Dating

Databases

GeoChron (EarthChem)

Software

MASS SPEC - Ar-Ar data reduction
ArArCalc - Ar-Ar data reduction
U-Pb_Redux - U-Pb Data Reduction
Schmitz and Schoene spreadsheet - U-Pb Data reduction
Tripoli - Mass spectrometer data handling
SQUID (SIMS U-Pb data acquisition and reduction)
Isoplot (calculation and plotting of isotopic data)
Iolite – used for LA-ICP-MS U-Pb data reduction
VizualAge – used for LA-ICP-MS U-Pb data reduction
UranOS – used for LA-ICP-MS U-Pb data reduction
U-Pb Age for R – used for LA-ICP-MS U-Pb data reduction
PyChron - Jake Ross, ArAr data acquisition and reduction
ArVert - Bruce Idleman - inverts thermochron data for T-t history
Ar inversion/MDD codes: Lovera, Zeitler, Lister...
Glitter

PepiAge
TrackKey
U of A LaserChron Excel macros

Low Temp thermochron (U/Th-He + Fission Track)

Databases

National geothermal data system (in development)

Software

HeFTy (low-T thermochronology thermal history modeling)
Helios ((U-Th)/He data reduction and age calculation)
TracKey - Istvan Dunkl, FT data reduction
Radial plotter - Pieter Vermeesch, visualization
FT data reduction program development in progress through EarthChem
Helioplot - Pieter Vermeesch, visualization, population deconvolution
QtQt - Gallagher, tT simulations
PECUBE/Glide - Braun, thermomechanical model
HEMP

Quaternary geochronology (cosmogenic, U-series, OSL)

Databases

Software

StalAge - stalagmite specific age model
Copra - stalagmite specific age model, linking proxy data to age models
CRONUS (Balco, Cosmogenic Radionuclide Calculator)
CosmoCALC (Vermeesch, cosmogenic)
Chloe (Fred Phillips, cosmogenic)
OxCal (Ramsey: Poisson depositional models)

Detrital geochronology

Databases

Software

Kernel density plotter - Vermeesch
MuDiSc - multidimensional scaling for matlab and r
BinomFit - thermochron data
BayesMix - Gallagher

Terrestrial Paleoclimate/paleobiology

Databases

Neotoma Paleoecology Database (and constituent databases)
Paleobiology Database

NOAA/NCDC Paleoclimatology Database
Modern taxonomic databases (e.g. Tropicos, Mammal Species of the World, WoRMS,…)
GenBank
CMIP database
National Soil Carbon Network, International Soil Carbon Network
PANGAEA

Software
Laboratory of Tree-Ring Research, University of Arizona repository
ITRDB Data Bank – tree ring repository

Surface process/ landscapes/ remote sensing
Databases
UNAVCO (GPS/geodesy data)
OpenTopography.org (database and inventory for LiDAR and topographic data)

Software
CSDMS (Community surface for landscape models) – repository for models and computer source
Cascade (numerical model)
CHILD (numerical model)

Radiocarbon
databases
14C Near Eastern Radiocarbon Context Database
Archaeological Site Index to Radiocarbon Dates
Canadian Archaeological Radiocarbon Database
New Zealand archaeological radiocarbon database
Online 14C databases
Oxford Radiocarbon Accelerator Unit
RADON – Radiokarbondaten online (European)
AUSTARCH ver. 3 - database for 14C and luminescence
INSTAAR radiocarbon date lists (not digitized but huge resource)
tDAR (the Digital Archaeological Record)

Software
IntCal - international calibration curve
BCal (online Bayesian radiocarbon calibration tool)
CalPal (Cologne Radiocarbon CALibration and PALaeoclimate Package)
WinCal25 (The Groningen Calibration Program)
CALIB 6.0 (radiocarbon calibration program)
CaliBomb
Metabase (laboratory management software)
OxCal

Fairbanks calibration program
Canadian Archaeological Radiocarbon Database (CARD)
Bacon/Clam/BLT
Bpeat – Age depth modeling

Sedimentology/stratigraphy/paleoclimate

Databases

SedDB (Columbia/Lamont series of databases)
Global Sand-Sea database Int Assn Aeolian Res at DRI
NOAA/NCDC - paleoclimate proxies, unstructured, standard repository, but not easily searchable
Gridded climate data (WorldClim, PRISM, HADCRU
Climate Data Guide -- an inventory of climate data with synopses of data
IODP databases - sediment data for astrochronology, age model info
NSIDC, NGDC, MGDC
PLIOMAX/EYEGLASS-- Sea level focused.
Janus (IODP)
LIMS – IODP
MagiC (magnetostratigraphic data)
CHRONOS/Neptune – foram and biostrat, timescale
NOAA, SEDIS, LIMBS, JANIS (paleo-oceanographic & magnetostratigraphic data)
PANGEA (European database w/ IODP support)

Software

LOWESS (seawater Sr)
Macrostrat

Astrochronology

Databases

Software

astro: An R package for Astrochronology, Steve Meyers
Analyseries
Laskar - astronomical solutions, widely access
SSA-MTM toolkit
K-Spectra – Commercial version SSA-MTM
Past - Time series analysis/Paleobiological analyses
Manfred Mudelsee codes - Time series analysis
Lorraine Lisiecki codes – Match and Autocomp software, astronomical stacks
Peter Huybers codes – Matlab scripts for time series analysis
Arand – Time series analysis

Visualization tools

Software
GeoMapApp
EarthObserver (GeoMapApp for mobile devices)
Google Earth
ArcGIS
UNAVCO viewers

Community communication/dialogue tools
Discipline-specific
EARTHTIME website
Noblegasnetwork
Neptune listserv
PlasmaChem
OnTrack Form (limited activity)
USeries.org (not active)

Software-specific
ROpenSci - open source development, R community
GitHub - online forum for code management + sharing
sourceforge,
vhub

Organizational/standardization/governance tools
SESAR – International Geologic Sample Numbers and related metadata
DOI registration services (e.g. DataCite)
NOSC
iupac for isotope compositions
ICS (internat comm. of stratigraphy) online tools, also at chronos.org

Physical sample archives
Terrestrial rock repositories - USPR (polar repository), Smithsonian
Meteorites - ASU, Smithsonian, American Mus Nat Hist, Brit Nat Hist Museum,
Antarctic meteorite collection, Southwest Meteorite lab, UNM Institute of Meteoritics
Core repositories - Lamont, IODP, OSU, FSU, LacCore (lacustrine), Houston, Rutgers, State Geo.
Surveys
OSU Antarctic rock repository
Sedis – Collections based

Appendix 3- Outcome 4: Data and cyber-capabilities required

Tools that feed data into Domain-Specific Data Systems

- Community-established protocols for level 0 data acquisition
- Community-vetted algorithms that construct higher level data products

- Community-consistent and reproducible open-source data reduction platforms with “one-click” data upload (e.g., U-Pb_Redux/Gechron)
- Unique sample identification that serves to discover and federate like data across data systems
- Scalable, nestable, sub-GPS spatial reference frames (e.g. “Spot” concept of Structure domain workshop)
- Standards, spikes, reference materials
- Technique development
- Decay constant calibrations
- Each (sub)discipline establish working groups to discuss protocols
- Open workflow in terms of data reduction
- Legacy data input (e.g. DeepDive, create incentives for community participation)

Tools that extract knowledge from Data Systems to be used within EarthCube

- Linkages and nesting with related and more distant databases (stratigraphic, paleo, structural, etc.)
- Dynamic scientific computing resources for domain-specific data analysis, assessment and visualization (e.g. bolting things like Isoplot, Cronus calculators, OxCal-type resources to the data system)
- Data visualizations linking data in space and time (e.g. Corewall, Geowall, Chronozoom)
- Interface data with GIS technology

Outreach and education

- Create the connection between EarthCube and federal agencies charged with hazards (e.g. USGS)
- Create the connection between EarthCube and industry (e.g. resources)
- Parallel portal to demonstrate utility of EarthCube to K-12 teachers, Congress and other policy makers (e.g. SERC)
- Resources for general public and for other scientists to understand and interpret geochronological data
- A geochronology data guide -- web based guide to where data is, what the data are, expert assessment of the data strengths and limitations.
- Means to invite outside communities to learn about and use our databases and software resources

EXECUTIVE SUMMARY: EARTHCUBE WORKSHOP RESULTS

(Kerstin Lehnert, Lamont-Doherty Earth Observatory of Columbia University: March 6 & 7, 2013)

Steering Committee: Chuck Connor, Elizabeth Cottrell, Rajdeep Dasgupta, Kerstin Lehnert, Abani Patra

Additional Report Authors: Maryjo Brouce, Adam Kent, Erik Hauri, Frances Jenner, Abani Patra, Julia Hammer, Christian Huber, Brendan McCormick, Amanda Clarke, Marion Le Voyer, Ken Rubin, Ben Andrews

Earth Cube Workshop Title: *Earthcube Domain End-User Workshop: Community-based Cyberinfrastructure for Petrology, Geochemistry, and Volcanology*

Introduction: More than 75 scientists, data managers, sample curators, and cyberinfrastructure specialists (on-site and remote participation) convened for 2 days at the National Museum of Natural History of the Smithsonian Institution in Washington, DC, to evaluate the status of cyberinfrastructure 'readiness' of petrology and, geochemistry, and articulate CI needs and requirements for these domains to contribute to the overall EarthCube architectural design phase. A late winter storm in the DC area obstructed travel and prevented more registered participants from attending on site.

Petrologists, geochemists, and volcanologists share the scientific interest in fundamental questions concerning the chemical and physical state of the Earth, the Moon and the other terrestrial planets; processes that have lead to physical and chemical differentiation and evolution of planetary interiors and environments through time; and the relationship between geologic processes and societal issues such as natural hazards and resource use. They generate data in the field by collecting samples or monitoring volcanic activities, in laboratories by performing chemical analyses or physical experiments, and by using these observational data to compute models.

The workshop identified important scientific drivers for advancing cyberinfrastructure in the domains, technical, data-related impediments to addressing scientific challenges, and resulted in a list of recommendations for next steps to realize the cyberinfrastructure vision for this community. With 28 science scenarios submitted in advance of the workshop and a record participation in the EarthCube Stakeholder survey, this community demonstrated a high level of engagement that continued throughout the discussions in plenary and breakout sessions.

SCIENCE ISSUES AND CHALLENGES

1. **KEY SCIENCE QUESTIONS or Important science drivers and challenges:** Participants identified

several high-priority science questions that will be the focus of and grand interdisciplinary efforts during the next 5-15 years:

- Understand the co-evolution of the geo- and biosphere
- Create a four-dimensional (space-time) description of the chemical and physical state of the Earth including the composition of and extent of fluxes between its major reservoirs -- core, mantle, crust, biosphere and hydrosphere.
- Understand the role of disequilibrium processes in the formation and evolution of planetary bodies.
- Integrate observations at volcanoes (e.g. seismic activity, ground deformation, emissions, magma chemistry and petrology, magma physical properties, plate tectonic parameters) in order to (1) forecast and mitigate natural hazards (2) understand and communicate volcanic impacts on society, and (3) to search for natural resources.
- Map the feedbacks between planetary evolution, plate tectonics, volcanic activity and climate on short and long timescales.
- Communicate the grand challenges of science to society.

2. **Current CHALLENGES to high-impact, interdisciplinary science:** Several categories of scientific challenges arose during the breakout sessions. Unless addressed, they will likely serve as significant impediments to conducting high impact, global scale research with a new Earth Cube infrastructure. These challenges can be broken into the following categories:

- Data challenges: Limitations in accessible data types, diversity and extent of temporal and spatial scales of existing data sets (mostly collected for regional studies and with different goals), variations in meta data standards or lack thereof, differences in data documentation in different scientific community. All of these issues manifest mismatch between marine and terrestrial data resources, which will need to be merged and standardized to address global scale variations and patterns in magmatic phenomena.
- Geosample Strategies: Most geologic terrains of interest to this end-user community do not have sufficient or even sample density through time and space. In addition, there is insufficient appreciation of this problem by many current geochemical/petrological practitioners and funding agencies.
- Sample Curation: Poor and uneven access and management of sample collections, incomplete sample tracking and linking of samples to analyses in the literature or databases, discoverability of existing samples.
- Knowledge: Lack of basic knowledge of limitations and uses of data and models across, within

and between disciplines.

- Interdisciplinary Conceptual Framework: Missing conceptual models, gaps in understanding, for instance of process rates, insufficient geochronometers over the full time range.
- Community: Barriers to collaboration, incomplete shared knowledge of data resources, expertise, specialized skills, toolsets, lab capabilities, etc. The community needs a “facebook” style networking site for geochemists and aligned scientists as a means to bridge this barrier.

3. **Technical Information / Issues / Challenges:** The workshop participants identified an initial set of desired capabilities (tools and databases) needed for pursuing key science questions:
   - Discoverability: Improved infrastructure (search engines, catalogs) is needed that facilitates discovery of data, samples, models, and management tools.
   - Interoperability: Software packages utilized during data acquisition should be transparent to data analysis and visualization softwares. Tools should support analytical thinking and numericisms.
   - Compliance: Data and metadata should be captured at the point of acquisition in a way that they can seamlessly be managed throughout their life cycle, including upload to repositories in order to satisfy data management requirements.
   - Format standards: Standards need to be established within the existent data repositories.
   - Sample tracking: Systems should be in place to promote spatial contextualization of analysis through sample registration, imagery, and links between samples (hand samples, thin sections, splits, etc.) and analytical data.
   - Archiving: Absent repositories, databases, and heterogeneous media should be identified and recovered.
   - Metadata: Ancillary contextual information such as science objectives, data provenance, and uncertainty estimates at each step in workflow needs to be included with the data.

4. **Community next steps:** Participants agreed on a number of steps that could be taken in the near future, many of which are ‘low-hanging fruit’.
   - Inventory: Create an online list of resources, a list of metadata for each technique/method, and an online list of science scenarios which fall under the “geochemistry/petrology” umbrella. These should be hosted on the EarthCube website and should be editable by registered users, but moderated and administered centrally. Individuals should start to populate the list of existing resources that they know of (data, samples, models, visualization tools, educational tools, expertise pool) and the list the metadata they use on a daily basis.
   - Participation: Get started! Pitch in! The following simple tasks should be taken on at the

individual level:

   - register on the Earthcube website and populate your own profile with picture/information
   - register your own samples (SESAR)
   - refer to the list of existing resource, start using them fully in your research and teaching, and encourage your students and colleagues to use them.

If everyone is doing a little bit of all of this, we will make HUGE progress rapidly! We need to encourage a change of mindset ("cultural shift") so that our community begins contributing to grassroots cyberinfrastructure.

- Social networking: Create a social network and expertise pool for early career and other scientists to exchange and build up the future EarthCube collaboration network/trust.
- Sample Curation: Consider samples as resources: we are not going to create a new HUGE sample repository, but just record and publicize which samples are where, who are the owners/ people responsible for each specimen, which analyses have already performed, how to access/borrow each specimen. A new library can be created, initially compiling existing online searchable sample catalogues and also encouraging collection managers and curators to digitize their own collection catalogues.
- Databases: Initiate work on immediate needs/low hanging fruits:
   - geospatial model of surface heat flux (e.g. Shapiro & Ritzwoller, 2006), plate boundary motions (e.g. Bird, 2003), crustal thickness (Bassin, 2000)
   - tephrachronology database
   - gas emission database
   - comprehensive, well standardized, error-documented volcanic rock composition database (marine and terrestrial)
   - getting historical data properly represented in EarthChem
   - GeoPRISMS as a venue to publish data (private and in grey literature)
- Community building: Self-organize in small working groups made up of people with similar science goals. Attempt to articulate specific science drivers and identify data formatting and cataloguing challenges. For example,
   - Forming an IAVCEI Commission of Explosive Volcanism working group to assess the need for a database of semi-quantitative data (video) of explosive eruptions and analogue experiments.

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/.

EXECUTIVE SUMMARY: EARTHCUBE WORKSHOP RESULTS

(PI: Anders Noren, University of Minnesota: February 4-6, 2013)
Steering Committee and Summary Authors: Anders Noren, Julie Brigham-Grette, Kerstin Lehnert,
Shanan Peters, Jack Williams, Emi Ito, Dave Anderson, Eric Grimm

EarthCube Workshop Title: Cyberinfrastructure for Paleogeoscience

Introduction

Forty-five participants gathered in Minneapolis to assess the current state and trajectory of the paleogeoscience domain, here defined as the community of scientists working to establish and synthesize paleorecords of Earth environment and biosphere. The defining characteristics of this broad domain are: 1) its focus on past earth and life processes; and 2) all scientific inferences in this domain are ultimately based on the collection of physical samples in the field, from which many kinds of geochemical, geobiological, and geophysical measurements are extracted. This domain includes (but is not limited to) scientists working on paleorecords from: cores drilled in the seafloor, lakebeds, peatlands, continental crust, glaciers or ice sheets, or trees; rock samples hammered from outcrops; fossil remains retrieved from various depositional environments; speleothems; corals; boreholes; packrat middens; etc. Participants' expertise spanned most of these disciplines, with heavy emphasis on the many subfields of paleoclimate and paleobiology. Outreach to the small proportion of communities not represented was accomplished during the four months of workshop preparation, through postings to relevant listservs, direct communications to community representatives, town hall gatherings at the GSA and AGU national meetings, and an online survey tailored to the community.

The main themes from workshop discussions are summarized below. Included as appendices to this summary are: a) a list of workshop participants and affiliations; b) full notes from all workshop breakout sessions; c) an inventory of current community cyberinfrastructure resources; d) the results from the online community survey.

SCIENCE ISSUES AND CHALLENGES

1. **Important science drivers and challenges:** Participants identified several high-priority science questions that will be the focus of interdisciplinary efforts during the next 5-15 years. Some of those common themes are described below.

Overarching theme: The History and Future of Life and Environment Interactions on Earth

- Establishment of a 4D framework for life and its environment on Earth. All other community priorities emerge from this primary objective. This framework will integrate across all time scales, regions, taxa, physical/geochemical properties, etc., and enable the ability to extract system state and rate of change at any spatiotemporal moment of interest.
- Determine climate/ocean/biosphere interactions during times of great change in climate and environment, including extinction events, periods of extreme warmth, and changes over decades to millions of years, including the present geologic transition. Develop detailed characterizations of these past events to inform predictions of future changes.
- Advance the capability to model the coupled carbon-climate Earth system, deriving the feedbacks, tipping points, and other processes from the paleo record, which is especially critical for deciphering the high magnitude/slow feedback mechanisms (e.g. ice sheet loss, deep ocean circulation) that climate models do not yet fully incorporate.
- Assimilate paleo observations into process-based Earth System Models to reconstruct Earth history (lat, long, elev, and time), developing a suite of products that facilitate research, inform policy and decision-making (carbon cycle, sustainability, hazards), and deepen the public understanding of environmental vulnerability.

- The participants recognized the value of recent efforts in characterizing many of the critical science drivers and challenges, including:
    - TRANSITIONS: http://www.sepm.org/CM_Files/ConfSumRpts/TRANSITIONSfinal.pdf
    - NROES: http://www.nap.edu/openbook.php?record_id=13236
    - DETELON: http://detelon.org
    - Conservation Biology Workshop Report: http://www.paleosoc.org/CP_Workshop_Report_Oct_2012.pdf
    - IGBP PAGES Report 57: http://www.pages-igbp.org
    - NRC 2011 Report: www.nap.edu/catalog/13111.html *[added by steering committee]*

Advanced cyberinfrastructure can enable the paleo community to reach these goals by 1) integrating small pieces of information scattered across the long tail (many small science projects), 2) refining sample ages and age uncertainty requisite to meet above challenges, and 3) facilitating a new era of vigorous collaboration across the many subdisciplines within and outside the paleogeosciences (e.g., hazards, paleomagnetics, tectonics, climate impacts, resource management, STEM education).

2 Current challenges to high-impact, interdisciplinary science:

Several themes emerged as consistent challenges faced within/across the paleorecord community.

Data/IT issues:

- Difficulty of discovery and vetting legacy data and dark data and their associated metadata; lack of funds for digitizing legacy data
- No consistent mechanisms for tracking changes in chronology or taxonomy (data product evolution)
- Lack of standards for age/time data
- Unstructured data
- Unawareness and/or underutilization of standards for data/metadata
- Inconsistent data formats
- Difficulty of importing/exporting data from databases
- Expense of IT development and maintenance
- Multiple databases for certain data types increases difficulty of finding data of interest

Social issues:

- Some disciplines do not have organized databases or sample repositories
- Barriers to data contribution (time, unease dealing with data formats and portal interfaces, perceived lack of incentive to contribute data, lack of citation for reused data)
- Perceptions of data ownership, personal investment in data and reluctance to share
- Lack of community organization

TECHNICAL INFORMATION/ISSUES/CHALLENGES

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

Breakout session 4 addressed the needs of the community in satisfying the science and educational objectives. The following summary attempts to represent the themes identified independently by two or more of the groups, organized roughly according to group reports. These items represent the minimum requirements needed to reach the overarching community goal stated above. NB: “existing resources” here means databases, museum digital and specimen holdings, models/model results, and data analysis tools/methods.

- Intuitive 4D access to all existing knowledge products and underlying data/metadata and methods used to generate those knowledge products
- Intuitive 4D mapping and visualization capability across data resources/model outputs
- Better access to and discovery of tools and methods to manipulate and analyze data of different common types (e.g., time series, stratigraphic position)
- Improved agreement upon standards and semantics for basic, widely-used data/methods, particularly for age/time representation
- System for determining and dynamically updating age models (and uncertainties) within and between existing resources and model results
- Improved user workflow and explicit reward system for data generators (e.g., acquisition and submission to databases/repositories
- Coupled earth-life system models that have good two-way, "live" integration with distributed data resources
- Increased awareness/utilization of existing resources within and outside of the paleo community and funding to sustain and improve these resources
- Improved metrics to evaluate success and contributions of existing efforts on a community and individual level; metrics to evaluate successes of new efforts
- System to identify gaps in existing data sets and prioritize/incentivize verification of contradictory information, as well as filling gaps with new records
- New educational capability that is built upon data and results drawn "live" from existing resources
- Support for long-term archiving and retrieval of digital data/tools and physical samples
- Need 4D visualization(s) for researchers (easy data comparison, discovery of gaps, etc,), scientists outside the paleo community, educators, policy-makers, and the public
- Legacy and dark data incorporation – noisy signal processing (sort out bad data, dropped data, sparse matrix data (missing images, geochem, geomorph; this type of tool is also fairly standard but it needs to be incorporated)

COMMUNITY NEXT STEPS

1. **List of what your community needs to do next to move forward how it can use EarthCube to achieve those goals:**

Develop an EarthCube RCN proposal with the goal of building an EarthCube community in Paleogeoscience.

- Working groups representing subdomains such as paleobiology, paleoclimatology, paleoceanography, curation, and cyberinfrastructure were established with initial membership to contribute to the content of the proposal
- Proposed activities will:
  - generate awareness of EarthCube and broaden participation within the paleo community, especially engaging early career scientists; consider iDigBio model of building working groups that focus on specific topics to increase participation
  - explore ways to advance the integration of the paleo community; reach out to other communities that have successfully transitioned from dispersed to integrated, and learn from them; develop incentives for participation
  - create a comprehensive inventory of existing data resources, models, and tools, and assess and evaluate them with support from computer and informatics scientists to determine optimal means of creating virtual connections between the resources

    - review existing data/metadata standards and coordinate with established interdisciplinary standards and publication/citation organizations to develop relevant community-specific standards, including ontologies and vocabularies (with particular emphasis on the urgent need to develop standards for time/age representation), and best practices for data citation and data publication
    - establish a broad initiative across EarthCube to describe and assess current methods of age representation in the geosciences and plan for a Building Block proposal to address this topic
- Community engagement will continue through contact on the EarthCube website, posting to listservs, webconferences, and AGU and GSA EarthCube sessions and town halls.

USE CASE EXAMPLES

1. Big Science Use Case: Earth-Life Transitions

We want to be able to fuse all forms of paleo evidence into a common 4D spatiotemporal framework, establishing what we empirically know about the past transitions in the Earth's system and its biota. This system makes it easy to find, search, and compare paleodata and to fuse these data with process-based models of the Earth-Life system. We move from incomplete and disconnected domains of knowledge into comprehensive and interlinked characterizations of the past Earth System and its biota. There is a seamless chain from primary observations to high-level data products and all information can be linked back to the original investigators. Data-derived products (e.g. past sea level, global temperature fields, species distributions, and rates of extinction) are dynamically updated as new information is generated, assimilated with earth system models.

Utility of Scenario: Such a framework would highlight data hotspots as well as baldspots for future data collection effort. A comprehensive framework would allow comparative studies that are the basis for testing causal hypotheses in historical science. The critical societal need is a model that can forecast Earth-Life system response at multiple spatiotemporal scales into the future. This model forecast is constrained by the data and knowledge collected by the entire paleo community over the past decades and the data-model fusion is advanced by state-of-the-art analytical and visualization tools.

2. Use Case: Empowering Individual Geoscientists

Now we take the prior example and invert it, to represent the perspective of an individual geoscientist.

Zoe is a paleolimnologist specializing in the reconstructions of salinity and temperature from fossil diatom assemblages extracted from lake sediments. She works with small research teams of other paleolimnologists, paleoecologists, geochemists, paleoclimatologists, each of whom specializes in the measurement and analysis of a particular kind of 'proxy', from various physical, geochemical, and biological sources (e.g. diatoms, ostracodes, stable isotopes, organic geochemical biomarkers). These data are of great interest to Earth System modelers seeking empirical constraints on their simulations of past transitions in the Earth-Life system (e.g. the Paleocene-Eocene Thermal Maximum, the Younger Dryas), and also to other paleolimnologists seeking to discern larger patterns from their individual time series. EarthCube has the potential to smooth every step in this scientific workflow.

Project Planning: During the initial planning stage, EarthCube allows Zoe to determine an optimal site, by confidently identifying 'bald spots' (places where no cores have been taken, the measurements from earlier work are no longer adequate to answer current questions, and/or the original core samples have been lost or destroyed). EarthCube also lets Zoe access archives of paleoclimatic model simulations and identify times or places of model ambiguity that would benefit from new observations.

Field Work: While Zoe is in the field, she tags each core collected with a unique digital object identifier and is able to digitally upload and link geospatial coordinates, field photos, and other field observations to relevant EarthCube repositories as they are collected in the field and in the lab. Zoe can easily share the data with colleagues and has the option to embargo sensitive portions of her data until they are published.
Lab Work and Project Management: Zoe's identifications of diatoms are checked against online reference libraries consisting of images of diatom species that are community-curated and synchronized with taxonomic databases. Visualization software makes it easy for Zoe to jointly plot and share her data with her team members during monthly teleconferences and for the entire team to set Zoe's data in the context of existing paleorecords.
Analysis. Next-generation software allows Zoe to use state-of-the-art statistical methods, e.g. to 1) build age models (needed to infer the time dimension) and 2) reconstruct past salinity and water temperature based on the environmental tolerances of the assemblages. Often these statistical methods will borrow strength from other previously collected datasets. Age models would use the most up-to-date timescales and standards and advanced estimates of uncertainty. This allows times-slices to be rapidly created using standardized data.
Synthesis: Once Zoe has finished her analyses, she can easily compare her data to other time series data stored in online public repositories and to the output from Earth System models that have been run previously for similar time periods.
Sharing: Upon publication, her diatom observations are made publicly available through EarthCube-affiliated federated repositories as are her age models and paleoenvironmental reconstructions. Her data are automatically incorporated in on-going larger-scale syntheses of paleoclimatic data and made available for subsequent data-model assimilations.

3. Key Concepts

- Seamless movement between data and models.
- Integrated Earth Systems models (interlinked atmosphere, ocean, biosphere) that are built iteratively to fit available data that is assimilated continuously from multiple data networks.

4. Key Needs:

- Assessing current informatics resources. Finding and sharing best practices.
- Global Access to Global Collections: establish repositories for all physical samples and the biological, geochemical and physical measurements made from those samples.
- Automated tools for finding 'dark data' and adding this data to repositories.
- Targeted Data Rescue campaigns for legacy data not in digital format or in obsolete, or difficult-to-use formats.
- Database linking: Improve connectivity among existing databases through adoption of common standards, establishment of standard and shared digital object identifiers, and/or shared semantic/ontology frameworks for linking between databases.
- Creation, enhancement, and sharing of workflow and data-management software designed for research teams ranging from a few scientists to large drilling campaigns.
- Formalization of "Level 0/Level 1/Level 2/Level 3" data products (i.e. ranging from the raw field and lab measurements to interpretations, global reconstructions, and other products; higher the level, more processed) and developing expert-guided workflow software that can dynamically update higher-level products.

Appendix A: Workshop Participant List

First Name	Last Name	Affiliation
David	Anderson	NOAA
Nicole	Anest	Columbia University/LDEO
Franco	Biondi	University of Nevada, Reno
Gabe	Bowen	University of Utah
Julie	Brigham-Grette	UMass Amherst
Marjorie	Chan	University of Utah
Mark	Chandler	Columbia University
Emilie	Dassié	Columbia University/LDEO
Edward	Davis	University of Oregon
Josh	Feinberg	University of Minnesota/IRM
Doug	Fils	Ocean Leadership
Russ	Graham	Penn State University
Eric	Grimm	Illinois State Museum
Ben	Hardt	USGS
Sonja	Hausmann	University of Arkansas
Sean	Higgins	Columbia University/LDEO
Brian	Huber	Smithsonian
Virginia	Iglesias	Montana State University
Randall	Irmis	University of Utah
Emi	Ito	University of Minnesota
Chris	Jenkins	University of Colorado/INSTAAR
Jim	Klaus	University of Miami
Kerstin	Lehnert	Columbia University/LDEO
Xiaoming	Liu	Carnegie Institution of Washington
Amy	Myrbo	University of Minnesota/LacCore
Charles	Nguyen	University of Minnesota
Anders	Noren	University of Minnesota/LacCore
Ryan	O'Grady	University of Minnesota/LacCore
Thomas	Olszewski	Texas A&M University
Shanan	Peters	University of Wisconsin-Madison
Surangi	Punyasena	University of Illinois
Frank	Rack	University of Nebraska/ANDRILL
Josh	Reed	ANDRILL
Alison	Smith	Kent State University
Dena	Smith	University of Colorado
Scott	St. George	University of Minnesota
Joe	Stoner	Oregon State University
Anne	Thessen	Data Conservancy
Michael	Tuite	Marine Biol. Lab/Data Conservancy
Amanda	Waite	University of Florida
John	Wehmiller	University of Delaware
Jessica	Whiteside	Brown University
Nancy	Wiegand	University of Wisconsin-Madison
Jack	Williams	University of Wisconsin-Madison

NSF Liaison

Lisa	Park-Boush	NSF

Appendix A: Workshop Participant List

First Name	**Last Name**	**Affiliation**
David	Anderson	NOAA
Nicole	Anest	Columbia University/LDEO
Franco	Biondi	University of Nevada, Reno
Gabe	Bowen	University of Utah
Julie	Brigham-Grette	UMass Amherst
Marjorie	Chan	University of Utah
Mark	Chandler	Columbia University
Emilie	Dassié	Columbia University/LDEO
Edward	Davis	University of Oregon
Josh	Feinberg	University of Minnesota/IRM
Doug	Fils	Ocean Leadership
Russ	Graham	Penn State University
Eric	Grimm	Illinois State Museum
Ben	Hardt	USGS
Sonja	Hausmann	University of Arkansas
Sean	Higgins	Columbia University/LDEO
Brian	Huber	Smithsonian
Virginia	Iglesias	Montana State University
Randall	Irmis	University of Utah
Emi	Ito	University of Minnesota
Chris	Jenkins	University of Colorado/INSTAAR
Jim	Klaus	University of Miami
Kerstin	Lehnert	Columbia University/LDEO
Xiaoming	Liu	Carnegie Institution of Washington
Amy	Myrbo	University of Minnesota/LacCore
Charles	Nguyen	University of Minnesota
Anders	Noren	University of Minnesota/LacCore
Ryan	O'Grady	University of Minnesota/LacCore
Thomas	Olszewski	Texas A&M University
Shanan	Peters	University of Wisconsin-Madison
Surangi	Punyasena	University of Illinois
Frank	Rack	University of Nebraska/ANDRILL
Josh	Reed	ANDRILL
Alison	Smith	Kent State University
Dena	Smith	University of Colorado
Scott	St. George	University of Minnesota
Joe	Stoner	Oregon State University
Anne	Thessen	Data Conservancy
Michael	Tuite	Marine Biol. Lab/Data Conservancy
Amanda	Waite	University of Florida
John	Wehmiller	University of Delaware
Jessica	Whiteside	Brown University
Nancy	Wiegand	University of Wisconsin-Madison
Jack	Williams	University of Wisconsin-Madison

NSF Liaison

Lisa	Park-Boush	NSF

Note: For the sake of length, appendices B and C, *Breakout Group Notes* and *Paleo Community Cyberinfrastructure Resource Inventory*, and the results of the *EarthCube Cyberinfrastructure for Paleogeoscience Community* Survey have been removed from this version the EarthCube Cyberinfrastructure for Paleogeoscience Workshop Report. These materials are included in the version of the workshop report located at: http://workspace.earthcube.org/content/earthcube-workshop-results-cyberinfrastructure-paleogeoscience.

EXECUTIVE SUMMARY: EARTHCUBE WORKSHOP RESULTS

Earth Cube Workshop Title: *Deep Seafloor Processes and Dynamics*

V.L. Ferrini (LDEO) and K. Rogers (CIW, RPI)

June 5-7, 2013

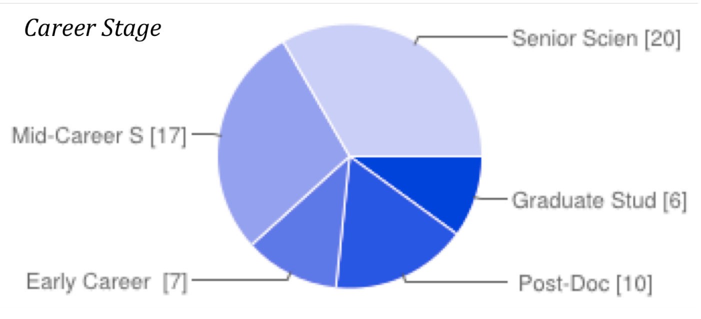
Introduction (field(s)/area(s) of interest and purpose, number of participants)**:**

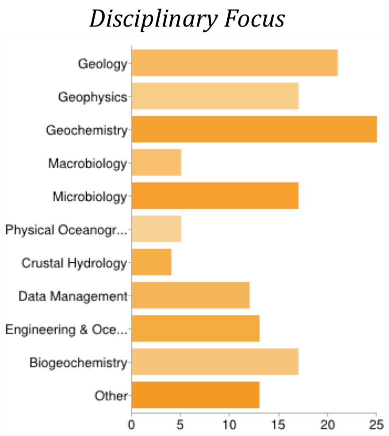
At the interface of Earth's interior and its external/surface environment lies the deep seafloor environment. The seafloor serves as the primary conduit for mass and heat transfer between sub-seabed and ocean systems, which operate on vastly different time and mass scales. Dynamics at this interface drive global (bio)geochemical elemental cycles, control global ocean chemistry, shape the surface atmospheric and climate system, and define the Earth's surface via tectonic processes. The seafloor-ocean interface also hosts some of the most diverse and extreme ecosystems in the biosphere, including deep-sea hydrothermal vents, cold seeps, mid-ocean ridges, deep-water coral ecosystems, ridge flanks, and plate margins, to name only a few. Research in deep seafloor processes spans a variety of disciplines as well - petrology, geology, geophysics, hydrogeology, aqueous geochemistry, micro- and macro-biology, ecology, and evolutionary biology - and the transformational science in deep seafloor systems occurs at the interface of these disciplines. True interdisciplinary research in deep seafloor dynamics requires mining and integration of large datasets from disparate disciplines and data integration and management are key components to the future success of interdisciplinary research in this field.

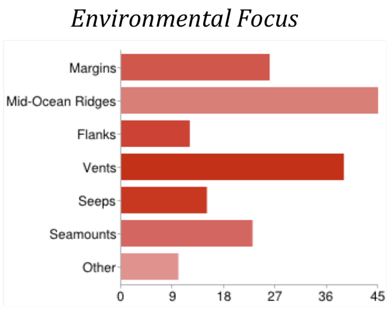
Scientists working in the deep seafloor environment comprise a model interdisciplinary end-user group that will benefit from the NSF EarthCube initiative. Integration of datasets generated by the deep seafloor research community could serve as a framework for analogous systems where integration of spatial and temporal cross-disciplinary data is crucial to the continued success of research efforts. The EarthCube End-User Domain Workshop for Deep Seafloor Processes and Dynamics targeted the major stakeholders in the deep seafloor research community and cyberinfrastructure specialists to chart the data integration and management needs into the EarthCube domain. As part of previous efforts to increase the participation of early career scientists in deep seafloor research, applications from graduate students, post docs and assistant professors and other early career scientists are especially encouraged.

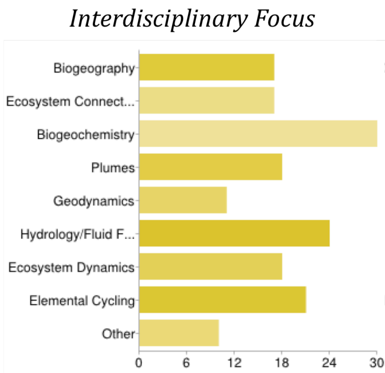
The total number of registrants for this workshop was 61, and an additional 2 remote, unregistered participants called in for portions of the workshop. Workshop participants were nearly evenly spread across career stages, with 20 Senior Scientists (16+ years experience), 17 Mid-Career Scientists (6-15 years experience), and 23 Early Career Scientists (< 5 years) including 6 Graduate Students & 10 Post-docs.

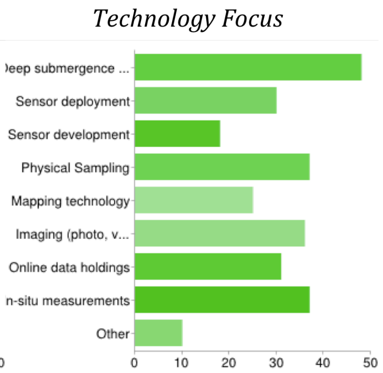
Workshop Participant Demographics


Career Stage


Disciplinary Focus


Environmental Focus


Interdisciplinary Focus


Technology Focus

The workshop included several invited speakers including technical and infrastructure perspectives as well as science perspectives:

- Technical and Infrastructure Perspectives
  - Eva Zanzerkia (NSF) -- EarthCube
  - Peter Fox (RPI) -- Geoinformatics and Cyberinfrastructure
  - Vicki Ferrini (LDEO) -- Services provided by the IEDA Data Facility
  - Giora Prioskurowski (UW) -- OOI
  - Dwight Coleman (URI) -- Deep Submergence Telepresence
- Science Perspectives
  - Scott White - (Univ. S. Carolina) -- Geology
  - Breea Govenar (RIC) - Macrobiology
  - Pete Girguis (Harvard) - Microbiology
  - Daniela DiIorio (Univ. Georgia) - Plume Modeling and Fluid Flow

The workshop consisted of several breakout groups and plenary sessions to address both the scientific and technical priorities and challenges within this community. The Science Drivers and Challenges were initially addressed by participants in Discipline-specific breakout groups, and Challenges to Interdisciplinary Science were then addressed by the Interdisciplinary breakout groups. Technical Issues and Challenges were addressed by the Disciplinary groups and Community Next Steps were developed in an open forum plenary session. The 5 disciplinary groups were Geochemistry, Microbiology, Macrobiology, Geology & Geophysics, Physical Oceanography and Crustal Hydrology, and the 5 interdisciplinary groups were Biogeochemistry, Biogeograpy, Geodynamics, Hydrology/Fluid Flow/Plumes, and Ecosystem Dynamics & Connectivity. Each breakout session was followed by a synthesizing open forum plenary session.

In addition to the invited speakers who gave perspectives on both scientific and technical issues, the workshop participants also participated in and heard the results of the Stakeholder alignment survey, given by Dr. Joel Cutcher-Gershenfeld of the University of Illinois. Dr. Cutcher-Gershenfeld's work was complemented by in-person interviews during the workshop. These were conducted by Charlie McElroy, who is a graduate student working with Dr. Cutcher-Gershenfeld. The scientific community's response to both aspects of this work were excellent, with significant participation and interest in the social aspects of our collaborative and interdisciplinary challenges.

SCIENCE ISSUES AND CHALLENGES

1. **IMPORTANT SCIENCE DRIVERS:** Participants identified several high-priority science questions that will be the focus of interdisciplinary efforts during the next 5-15 years.
   - What are the geological/geochemical/physiological/energetic limits of life? What are the boundaries between biological and abiotic control of chemical reactions? How does geochemistry influence microbiology and vice versa? How do we incorporate microbial data into large-scale (global) quantitative geochemical models? How does bioenergetics influence food web dynamics, productivity, energy transfer and nutrient cycles and transform elemental pools between ecosystem compartments? What is the biogeographic, functional and structural distribution of microorganisms and what are the environmental parameters that most influence these distributions? Can these environmental parameters be used as indicators of ecosystem structure and vice versa? How do we define and interpret biomarkers (e.g., paleomicrobiology)? What are the scales of biological responses to disturbance, both natural and anthropogenic and how are these responses reflected in ecosystem connectivity, the relatedness of organisms? Can genetic tools be used to track ecosystem responses to environmental parameters, including adaptation and evolution?
   - What is the architecture of the oceanic lithosphere (including magma processes), and what happens to the plate as it ages from spreading center to subduction zone, as a function of spreading rate, environmental variability, variable crustal architecture? How does plate maturation impact subduction, and what controls the size and cycles of earthquakes in subduction zones? What role does the magma lens play in helping control tectonics/seafloor morphology? Do hot spot/ridge interactions influence the development of oceanic core complexes? What controls the origin, distribution, evolution, and morphology of seafloor features (e.g. seamounts, sulfide mounds), and what is the relationship between these processes/environments on biological communities and mineral resources?
   - What is the role of the deep ocean and subsurface in obtaining a 4D (spatial and temporal) understanding of global chemical and biological reservoirs, fluxes, and energy transfer? Such a perspective would allow us to address such transformative questions as: How does Earth regulate atmospheric CO_2? What are the effects of deep sea biogeochemical processes on modern/ancient global atmospheric chemistry (C,O,S)? What are the relative contributions of biotic and abiotic deep ocean processes to global biogeochemical cycling? How can microbial data be incorporated into large-scale (global) quantitative geochemical models? What are the processes associated with serpentinization, including its diversity, range of environments, and consequences on global elemental cycles? How does the carbonation of peridotites affect global elemental cycles?
   - How do fluids in the subseafloor link thermal, tectonic, seismic, chemical and biological processes in a variety of deep-sea environments? What is the temporal evolution, extent and geometry of fluid flow within oceanic crust? What are the feedbacks between flow and geochemical and geophysical processes? How high within the water column do the fluids go? How does fluid flow effect the transfer of nutrients, energy and heat into habitable zones and what is the role of fluid flow is establishing geochemical gradients and (micro)niches of habitability within the crust?

2. CURRENT CHALLENGES TO HIGH-IMPACT, INTERDISCIPLINARY SCIENCE:

Several themes emerged as consistent challenges faced within/across the involved discipline(s)

- Data integration challenges
    - Communication between more disparate disciplines is lacking in large part because both scientific and funding links are tenuous, but also because there is little history of interaction across these disciplines (e.g., physical oceanographers at biogeography discussions). A key part of future success in interdisciplinary deep ocean studies is encouraging and facilitating communication between disciplines. Progress in this area will subsequently help to overcome the integration of data sets and discipline approaches that are described below.
    - Cross-disciplinary data integration is extremely challenging and true co-registered interdisciplinary data sets are the exception rather than the norm. These challenges stem from issues both at the data collection/management level and with data analysis. For example, data from different relevant disciplines (e.g., biological, geochemical, physical) are not often linked and even the same categories of data are often not comparable in key ways (for example, different molecular samples are processed in different ways and subject to different biases). Few cross-disciplinary data sets exist and deficiencies in acquisition protocols, data quality, and sample metadata make it nearly impossible to link data sets from different disciplines collected on different expeditions. Furthermore, advances in modeling and data analysis techniques are needed to improve cross-disciplinary data integration. For example, merging chemical models with physical or transport models is a science still in its infancy, and new kinds of modeling techniques are needed to integrate heterogeneous data and address interdisciplinary science questions. Within the data management domain, there is both a desperate need for more data in all disciplines and the foreboding challenge of developing tools and platforms (e.g. cloud computing) to handle ever-increasing data volumes and to make data interpolations.
    - To what extent can approaches from one discipline be applied to transform research approaches in other deep sea disciplines? There is significant room for cross pollination of research approaches across disciplines and such activities could be facilitated by categorizing such activities within the scope of broader impacts. In essence, how can you be someone else's broader impact?

- Data acquisition/completeness - particularly with respect to co-registration, and spatial/temporal variability
    - Because deep sea ecosystems are particularly closely linked to geochemical cycles (primary productivity is primarily chemosynthetic and reliant on geochemical fluxes and gradients) there is a need to make spatially and temporally co-registered chemical and biological sampling the norm rather than the exception in deep sea ecosystems. Acquisition of co-registered data is a challenge to current and future deep sea scientists, however access and integration of co-registered data as well as the resolution of legacy disciplinary data into pseudo-co-registered data sets is a challenge that can be addressed by the EarthCurbe Initiative and subsequent data analysis and management tools.
    - The spatial and temporal scales of data collection are very different across disciplines, making interdisciplinary data integration challenging, and many more co-registered, interdisciplinary data sets, collected on comparable scales are needed. For the current data sets, biological occurrence data (in the ocean) are inherently patchy in time and space. Furthermore, many aspects of biogeography are not captured by taxonomic data (such as habitability and energy flows). Additionally, integrating biological data sets

(e.g. spanning the ecosystem from microbe to macrobe) and scaling data sets properly (e.g. measuring specific populations vs. entire communities), could allow us to determine the extent to which specific (keystone) populations drive ecosystem function.

- o Understanding ecosystem dynamics requires both the discovery of the required data and synthetic analysis. Therefore the role and challenge of network analysis is to find what you *aren't* looking for that is important – e.g. what data are missing that will make the ecosystem analysis much more robust? Furthermore, adaptive ecosystem behaviors, emergent behaviors, disturbance factors are all challenges to understanding ecosystem function and to developing cyber infrastructure for modeling. Can we develop multi-dimensional datasets to reflect entire ecosystem function?

TECHNICAL INFORMATION/ISSUES/CHALLENGES

1. **Training and Awareness** – Many members of this community recognize that they are supported by existing data management efforts, and clearly stated that they do not want EarthCube to "reinvent the wheel". That said, there is insufficient awareness of and access to existing tools and infrastructure - including data contribution and data discovery tools, open source software, visualization tools, and data analysis systems.
   - New tools need to be developed to improve both data management and data analysis. Particularly, there is a lack of tools that lessen the "burden" of data management and could be embedded in our scientific and daily workflows. New tools that allow for easier and earlier integration of data management activities within the workflow are essential for future data acquisition.
   - There is a large personnel gap in the community between data producers and data managers that could be resolved by facilitating training within the community to lower barriers to available tools and resources.
   - There is significant and well-founded concern that the community lacks sufficient resources for data preparation and that those efforts are not sufficiently recognized and rewarded. While infrastructure for citing data and has been established within several data systems (e.g. Data DOIs), nearly all professional citations continue to be focused exclusively on publications. Much of the hard work of data acquisition, data management, metadata production, data integration is largely unrewarded, lowering the incentive for proper data acquisition and curation and increasing the gap between data scientists and discipline scientists.
2. **Data comparison and integration** -- Datasets are often not fully comparable *because:*
   - Metadata are incomplete and inconsistent;
   - Navigational precision is problematic across deep submergence vehicles. It is essential that exact locations (x, y, z, t) are precisely identified for each sample;
   - Foci differ from project to project. Improving mechanisms for pre-expeditionary communication and developing tools to enhance collaboration (either at particular sites or for particular types of sampling projects) would maximize project utility and drastically increase funding efficiency;
   - Data formats and entry vary from project to project. This can be resolved with either format standardization or, preferably, algorithms that identify and correct for variation in format;
   - There is a lack (or a lack of awareness) of standardized methodologies to document sampling conditions, e.g., consistent definition of time stamps and locations for samples and measurements.
   - Data quality is poorly documented making data use outside the original research group and integration of disparate data sets inconsistent.

3. <u>Desired Tools</u>

- Collaborative Tools
  - Tools are needed to facilitate real-time collaboration before, during, and after cruises. These include live ship-to-shore feeds that enhance shore-based participation in sample collection and real-time data analysis. Thus, expedition goals could be dynamic and responsive to real-time data analysis. Furthermore the use of Ancillary Project Letters (APLs) or RAPID-type funding models would allow for interested parties (this could focus on early career scientists) to join expeditions (in person or remotely) to collect co-registered or associated data/samples, thus increasing expedition efficiency. This is important for field-going scientists and modelers alike. This would also lower the barrier for early career scientists to undertake sea-going research by allowing for smaller projects to be funded and completed prior to pursuing larger expedition funding.
  - Mechanisms to better communicate caveats and built in assumptions necessary for interpreting data and models. Models, especially, need to continue to be linked to scientific expertise.
  - "Alert" system that will notify the user of a new data submission of interest. This could be developed to include not only data acquisition updates, but also self-populating personal databases and subsequent data analysis. For example, if one were interested in a particular metabolic functional gene in hydrothermal environments, a search/analysis/model algorithm could regularly self-update and new function gene trees would be the product for the end user. This goes beyond data discovery, but also automates data analysis, allowing scientists to focus on data interpretation.
  - Experimental design, communication/ cooperation with various deep sea and related scientific communities
- Data Documentation Tools
  - There is a lack of tools (desktop, tablets, in the field (ships, ROVs etc)) that facilitate data documentation and capturing metadata that can be used broadly by our community. This is a critical gap that needs to be filled if we are to effectively and efficiently feed content into EarthCube.
  - We also need improved and expanded metadata and standardized metadata templates that easily identify units and commonalities (e.g. when, where (projection, coordinates), how (methods of collection, analysis), experimental design). Furthermore, we need to develop easy tools and simple guidelines for easily capturing metadata contemporaneously at the time of data acquisition.
  - Data quality is inconsistent - EarthCube should include consistent and rigorous mechanisms for objectively documenting and evaluating data quality.
- Visualization and modeling tools
  - Many existing tools require extensive training for effective use or are incomplete. This not only inhibits usage across our community, but also limits our ability to analyze legacy data or integrate and analyze disparate data sets.
  - EarthCube should include a clear and well-organized user interface with a well-documented set of modeling and visualization tools (with training documents) that can be improved or extended in modular form.
  - We need more data integration tools, including tools that easily allow you to merge cross-disciplinary data (different data types) and tools that allow users to look at multiple data sets on a global scale. One idea was: "EarthClip" (*J. Smith*) - Integrated digital (desktop) guidance to help you discover data, contribute data, comment on data quality, etc. (e.g. *"You may also be interested in…"*).

  - Easily accessible interface for using open source tools, without requiring installation on individual computers – cloud based, web page, all encompassing application.
  - Tools needed for interactive figures (3-D) for both processed and raw data.
- Current data sets are enormous and the volume and quantity of data is only going to increase (e.g. HD video is becoming the norm, acoustic datasets, and someday (soon) biologists will be sequencing entire genomes for every organism in a sample). Moving these data sets will be (and is now) an enormous challenge and current solutions are rather antiquated (e.g. we currently ship large hard drives around the globe in order to share data and collaborate on interdisciplinary projects). We need to transition to cloud-based platforms that allow analyses in the cloud with systems that are connected with ultra high bandwidth networks.

4. **Data Curation and Access Issues/Challenges**
   - Relational databases that discern both user interest and intent from search parameters are now common in ecommerce, and could be applied to scientific data searches. For example, when you search for a spatula on Amazon, it shows you a bunch of other spatulas that other users also looked at. Is there a way to have Earthcube know or learn from users about connections between datasets in order to improve data discovery?
   - Access to legacy data is important but is often difficult - EarthCube should include legacy data and/or clear links to legacy data, including ways to objectively evaluate the quality of legacy data. Incorporating legacy data into EarthCube is essential for maximizing its impact in the deep sea science community, however this community will only buy into this platform if there is guaranteed longevity.
   - Lost data sets as well as data sets that don't get pushed into the public domain are not uncommon. We as a community need to continue to be vigilant about data compliance. Can Earthcube make it easier to find and upload data into various databases? Can it be a two way street? Tools that lower the barrier between publications and data upload and curation to data repositories are essential in order to minimize lost data sets and ensure compliance with funding agency requirements for data management.
   - Reducing barriers to access include cross-directorate, cross-agency data linkages ("Data without borders"). This includes NIH-NSF cross communication, potentially combining geological data with 'omics data. Public and private as well as national and international agencies (e.g., ONR, Schmidt, Moore, NOAA, IODP, etc.) support deep sea data acquisition, making data multi-jurisdiction but there is no jurisdiction to the seafloor.
   - Broad-based, interdisciplinary seafloor models and data sets need to be integrated with surface and coastal models, ideally by incorporating all of these in the EarthCube platform.

COMMUNITY NEXT STEPS

1. **List of what your community needs to do next to move forward; how it can use EarthCube to achieve those goals:**

   - We recognize that much of our community is served by existing data management efforts, and recognize that EarthCube should build off these, rather than reinvent them. However, barriers still exist, and we need training to ensure that we can take advantage of existing resources, and to ensure that data are documented and curated accurately and efficiently.
   - As a community, we see very cost-effective rapid solutions to a number of problems that create data management obstacles, but we are unsure what mechanisms might provide funding to address some of the smaller data problems that confront us. While we recognize that there are funding opportunites in EarthCube, it is not clear if any opportunities exist to obtain funding for community-specific projects that could facilitate

inputting data/metadata into the paradigm (e.g. NDSF's Jason Virtual Van upgrades would benefit from cyberinfrastructure input).

- A small subgroup of workshop participants will explore an RCN and/or workshop proposal focused on documenting expedition-based needs. The goal of this effort is to facilitate community consensus to prioritize the needs for improving existing resources for documenting deep submergence field programs - specifically the Jason Virtual Van and Alvin Frame Grabber.
- Tackle Education, Training & Best Practices - PIs, Graduate Students and post-docs need training on how to use available tools for data management, access etc. This can be in coordination with existing data groups that serve our community (e.g. IEDA). NSF should support this training effort. The DEep Submergence Science Committee (DESSC) members who participated in the workshop will pursue this in the context of ongoing early career training efforts.
- Also discussed the concept of a "data wrangler" participating in field programs who is responsible for handling data, and can facilitate contemporaneous data documentation. The role of the data scientist who sits at the intersection of domain science and geoinformatics is rising, but resources are necessary to ensure good data management practices.

EARLY CAREER FEEDBACK - Early career participants conducted a small break-out session of their own to articulate their message to NSF and their perspective on EarthCube:

The dominant concern of our early career participants is related to funding and the bleak job market in academic science research. While they are enthusiastic about their research and the possibilities that EarthCube may enable for them, they are very concerned about career longevity and their ability to pursue cutting edge science in the academic environment.

They also suggested several actionable items we can strive for as a community that would better prepare them for doing better science in a data-enabled world including:

- Small grants for early career scientists to collaborate outside of their institutions.
- Enhanced training opportunities: For example, a data/computer literacy workshop would be broadly useful to early career scientists, and to the deep sea community as a whole.
- Encourage current PIs and mentors to include data/computer literacy into graduate curricula

NSF ACTION ITEMS

The group identified several action items for NSF that would immediately impact this group's ability to not only contribute to, but also be prepared for EarthCube implementation. While we recognize that a majority of the current funding opportunities in EarthCube are focused on developing a cyberinfrastructure to accommodate scientists across the Earth Science disciplines, we believe there are several issues within the deep sea science community that we need to address internally in order for our community to be fully prepared and part of the driving force behind the EarthCube Initiative.

- Develop a RAPID/EAGR-sized funding program (e.g. $50K/award) for discipline/community-specific projects that improve community resources so that they can be incorporated into the EarthCube Initiative. Examples of such community-specific projects include: (i) incorporating legacy data into current data repositories, this could also include rescue of (almost) lost data; (ii) improving data management tools for deep submergence assets (Alvin Frame Grabber, Jason Virtual Van, etc.); (iii) pilot programs for cross-disciplinary scientists to work with data scientists to integrate current data sets and develop small-scale, discipline specific tools that could later be incorporated into and expanded into EarthCube.

- Support data literacy workshops and programs that target both early career and senior scientists. Encouraging young scientists to be not only become data literate, but also to increase their marketability to fill a developing need for discipline trained data scientists within the deep sea community. Furthermore, senior scientists with expedition level responsibilities (e.g. Chief Scientists) need resources and training in data management and curation so that these activities are incorporated into expedition planning and become are integrated early in at-sea work flows.
- Support the incorporation of discipline data scientists into deep sea expeditions. We foresee this would be a multi-tiered program with both support for current data scientists to be integrated into the science expedition team and with support for training of expedition group members in data resource management during pre-expedition planning. We believe it is essential that every science party has a dedicated data scientist to facilitate shipboard data management and enhance data acquisition and documentation, which will serve both the immediate science expedition, and subsequent data users.

EXECUTIVE SUMMARY: EARTHCUBE WORKSHOP RESULTS

Earth Cube Workshop Title: Developing A Community Vision Of Cyberinfrastructure Needs For Coral Reef Systems Science

Convenors: Dr. Ruth D. Gates (Researcher, HIMB, UHM) and Dr. Mark Schildhauer (Director of Computing, NCEAS, UCSB)

Organizing Committee: Dr. Megan J. Donahue (Associate Researcher, HIMB, UH Manoa), Dr. Peter J. Edmunds (Professor, CSUN), Dr. Erik C. Franklin (Assistant Researcher, HIMB, UH Manoa) and Dr. Hollie M. Putnam (Assistant Researcher, HIMB, UH Manoa)

Workshops: 1. September 17-18, 2013, University of Hawaiʻi: Hawaiʻi Institute of Marine Biology (HIMB) and 2. October 23-24, University of California: Santa Barbara – National Center for Ecological Analysis and Synthesis (NCEAS)

INTRODUCTION: The purpose of the two workshops for the ***Developing a Community Vision of Cyberinfrastructure Needs for Coral Reef Systems Science*** project was to gather input from end users and data generators on the role that cyber-enabled data tools can play in addressing the key science drivers and grand challenges in the field, and in enhancing the value, scope and impact of coral reef systems science. A total of 53 participants, representing a broad geographic range of the U.S. academic coral reef research community, attended one or both workshops (in person or virtually). The coral reef scientific community focuses on a critically important and threatened ecosystem and is extremely diverse from a disciplinary perspective, crossing the boundaries of biological, physical and chemical oceanography, climate science, remote sensing, modeling and engineering. Research in the field spans genomics and ecosystem science, and data generated by these activities crosses broad biological, temporal and spatial scales.

SCIENCE ISSUES AND CHALLENGES

1. Key Science Drivers/ Questions in Coral Reef Science: Participants identified several high-priority science questions that will be the focus of interdisciplinary efforts during the next 5-15 years

- What processes are relevant to understanding the biological responses of coral reefs to biotic and abiotic drivers across temporal and spatial scales?
- What are the mechanisms of coral reef adaptation and acclimatization to climate change?
- How does symbiosis influence the biology and ecology of coral reef organisms?
- How does the abundance and diversity of coral reef organisms influence community resilience at local, regional, and global scales?
- How will invasive species, disease, and parasites disrupt coral reef ecosystem structure and function?

2. Grand Challenges in Coral Reef Science: Several themes emerged as consistent challenges faced within/across the involved disciplines.

- Data utilization and accessibility for automated processing, standardization and measurement was identified as the highest priority, and includes data cohesion across spatial and temporal scales, as well as disciplines, and application and access to this data, from omics to ecosystem modeling.

- Rapid developments in bioinformatics and –omics sciences provide new tools to address taxonomic, genetic, ecological, and evolutionary questions but there is a great need to develop methodologies to efficiently utilize these tools within the coral reef science community
- There should be improved training opportunities in communicating science to peers outside the field as well as to better inform policy and educate students and the public.

TECHNICAL INFORMATION/ISSUES/CHALLE NGES

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

There were four distinct classes of community needs: (1) Databases and Portals, (2) Data Processing, Modeling and Visualization, (3) Education and Training, and (4) Internationalization.

(1) Databases and Portals:

- Desirable features include standardization (formats, collections, and representation), richness of metadata, and built on existing efforts with tools to create and query data across repositories that includes standard reference keywords, DOIs, and appropriate credit for data provider. Data integration should support imaging, sequence, environmental sensor data, and local observational data and delivery of streaming real time data from sensors networks. Quality of data and metadata is critical since web applications and interfaces may involve with time.
- Needs to include links/connections to existing resources through a curated data portal that could cluster databases/data by types. The system should allow grass roots contributions through user data entry as well as information filtering and discipline or research theme by sub-setting. Digital tools should also provide a dynamic, collaborative workspace for a variety of sub-disciplines (bioinformatics, ecological studies, genomics, mathematical biology, programming, etc.). Identifying a funding strategy for sustainable data curation is critical.
- There are many “dark data” or challenging data types (such as imagery or sequence data) that could be better managed and harvested from unpublished studies, secondary reports, desk drawers, personal collections, original data from earlier publications not archived that require a various types of standards and integration methods. Improved methods to deal with these data types may arise from industry-academic technology sharing translated to the coral reef research community.

(2) Data Processing, Modeling, Visualization:

A system of data processing pipelines for bioinformatics/omics data for computationally intensive analysis tasks is critically needed especially those that take advantage of HPC resources (XSEDE). The KEPLER scientific workflow system is a potential tool to utilize. These approaches would include traditional statistical modeling approaches, machine learning, and geospatial analysis and multimedia analysis for image/video/audio analysis and information extraction.

- We could visualize disparate data (space/time) with “easy to use” software tools (*vs.* immersive environments) that support online visualization simulations with user-directed parameters. Google Earth is a reasonable model for the portal interface. The visualizations can be used to communicate directly with public through interactive and applied community engagement. Maintain data and software version control will tools such as GIT or Mercurial.
- Improved software tools (such as API’s: application programming interfaces) for linking ecological and -omics software packages are needed with open standards that facilitate coupling through modular-based software or middleware to connect processes; better interfaces for communication among software models; Free Open Source Software (FOSS); Glue code and provenance: automated metadata extraction/provenance from digital objects and (e.g., OpenDAP, HDF (file format): geodata, temporal metadata; Integration/ alignment: scale; measurement equivalence, reward coders and nurture new type of coral reef scientist/hacker
- Develop a coral reef simulation system that merges model components (forecast, climate change), is applicable to many locations; 3D; modular, with case studies; the WRF is an example (Weather Research and Forecasting)

(3) Education and Training (human resources/workforce)

- The support of a cyberinfrastructure tactical team to support training, scientific programming, and database administration would help facilitate many of the data analysis, education, and training needs. The coral reef domain aware CI team could rotate between thematic-based resources and help with challenging projects to achieve scientific end products that non-CI researchers would have difficulty creating alone. This role may also be fulfilled by a “Campus Champions” such as a grad student or similar to connect geoscientists and computer scientists.
- The coral reef community would support web-based workshops on portals for data; data integration; data management and mechanisms for local group participation (e.g., web-based). Various topics for education programs were proposed including programming, data management; online repositories/versioning; imaging technologies/analysis; tools for temporal/spatial scaling; communication with managers; promotion of cross-disciplinary training and research.
- University programs in coral reef sciences may institute computer programming requirements in curricula or develop and offer a Certificate Program in Coral Reef Informatics to encourage cross-disciplinary training (between geoscience and computer science)

(4) Internationalization

- Need to improve access and use of international data; issues include need to share data to improve our understanding of reef globally to promote international cooperation on global scale reef studies. We would like to connect people, institutions, government management to democratize research and get many contributors to interpret science and results.

COMMUNITY NEXT STEPS

- Maintain momentum and communicate
- Develop a formal network to facilitate the interface with one another and with EarthCube.
- Respond to solicitations for input from EarthCube such as the call for Research Coordinated Network proposals and Geoscience community activities
- Identify existing resources
- Summarize, visualize and communicate results to broader coral reef community
- Solicit feedback from the community on the proposed coral reef science scenarios and other workshop outputs
- Promote community buy-in by moving from “talking activities” to “implementation activities”
- Link EarthCube activities to ongoing activities, e.g., EPSCoR and other funding program
- Link to software infrastructure funding programs
- Identify possible partnerships for proposals
- Reach across EarthCube user groups to identify common needs and initiatives
- Conduct metrics on progress on recommendations, e.g., resource lists, status of specific workshop suggestions in 1 year, 2 year, etc.

EARLY CAREER STRATEGIC VISIONING WORKSHOP

Workshop Results and Executive Summary

October 16-17, 2012

"Over the next decade, the geosciences community commits to developing a framework to understand and predict responses of the Earth as a system—from the space-atmosphere boundary to the core, including the influences of humans and ecosystems."

GEO Vision report of NSF Geoscience Directorate Advisory Committee, 2009

Support for this workshop was provided by the NSF under OCI grant1256163, "Envisioning Success: A Workshop for Next Generation EarthCube Scholars and Scientists," for which we are most grateful. The session was held at the Carnegie Institution for Science and the wonderful hospitality and excellent support was deeply appreciated.

NSF EarthCube Workshop Results

Earth Cube Workshop Title and Date:

EarthCube Early Career Strategic Visioning Workshop

October 16-17, 2012

Co-Leaders and Institutions:

Joel Cutcher-Gershenfeld, University of Illinois at Urbana-Champaign

Steve Diggs, Scripps Institution of Oceanography

Yolanda Gil, University of Southern California

Bob Hazen, Carnegie Institution for Science

Danie Kinkade, Woods Hole Oceanographic Institution

Introduction (fields/areas of interest and purpose, number of participants):

Sixty-eight early career geoscientists and cyber/computer scientists gathered with five instructors and eight additional featured speakers (for a total of 80 participants) at the Carnegie Institution for Science on October 16-17, 2012 to construct a shared vision for success with respect to the cyberinfrastructure needed to support the next generation of earth science research. The participants were mostly assistant professors, but post docs, doctoral students, and a few others engaged in a highly interactive process of mapping their own career aspirations and considering how a robust cyber infrastructure might enable them to tackle high impact research questions and deliver education in new ways. All of the individuals invited to the workshop were selected based on their being seen as emerging leaders in their respective domains, which included the following domains at the NSF: AGS, BIO, CISE, EAR, HER, ENG, HER, OCE, OCI, OPP, and SBE.

Motivating the workshop was the new NSF initiative entitled "EarthCube," which is a ten-year initiative designed to create a knowledge management system and infrastructure that integrates all geosciences data in an open, transparent and inclusive manner. The overarching motivation was to understand how the research and educational trajectories of next generation leaders in the geosciences, computer sciences, and other relevant fields would map onto the future direction and potential for EarthCube. In particular, four goals for the workshop were identified:

- Map EarthCube onto Career Trajectories
- Contribute to EarthCube Vision
- Inform EarthCube Governance
- Enable Networking and Professional Development

Science Issues and Challenges:

1. Important science drivers and challenges:

- Participants were all motivated by "grand challenge" geoscience questions concerning global climate change, weather prediction, and other such challenges.
- This was a highly diverse set of participants, spanning the following geoscience domains (each with distinctive science drivers/challenges):
    - Atmospheric and geospatial sciences (anthropogenic aerosols, climate modeling, earth system science, land use, paleoclimate modeling, space science)
    - Earth Science (biochemistry, carbon cycling, climate change, climate modeling, earth system modeling, earthquakes, geochemistry, geochronology, geodynamics, geoinformatics, geology, geomorphology, geophysics, hydrology, igneous processes, metamorphic petrology, mountain environments, rivers, seismology, tectonics, water cycle)
    - Ocean science (biogeochemical cycling, chemical oceanography, climate change, coastal fluid dynamics, fluid mechanics, geochemistry, magmatic systems, microbiology, ocean acidification, petrology, physical oceanography, remote sensing)
    - Polar science (Antarctic ecology, carbon cycle, climate change, geochemistry, glaciers, ice, ice-ocean interface, meteorology, permafrost, sea ice)
- Additional participants were from cyber or computer science, social science, and other domains including:
    - Computer science (cognition, machine learning, software)
    - Cyberinfrastructure (algorithms, big data, bioinformatics, climate informatics, cyber data management, data mining, GIS, hydroinformatics, lexical representation, semantics, spatial/temporal data, special databases)
    - Education (disability, disasters, geology, soil and water)
    - Engineering (environmental nanotechnology, low temp geochemistry)
    - Social science (governance, stakeholder visualization, trust)

2. Current challenges to high-impact, interdisciplinary science:

- Institutional barriers to interdisciplinary science, particularly the tenure process in universities.
- Resources and credit for sharing data, tools, models, and software
- Connecting interdisciplinary research with interdisciplinary education
- Not being limited to "brute force" accumulation of interdisciplinary data, particularly where the "Z' axis for geochronology is needed.

Technical Information/Issues/Challenges:

1. Desired tools, databases, etc. needed for pursuing key science questions with brief elaboration:
    - One-stop shopping for improved access to data, ease of sharing data, with standardization, and ease in citing data – a "closed circle" from data production, use, review, and publication
    - Better funding of data storage options, with aligned ontologies, able to "keep up" with "big data," and capture of legacy and archival data
    - Minimizing time collating data and maximizing time doing science
    - "Hindcast" and predictive modeling capabilities
    - International access and access to the general public
    - Achieving the multi-disciplinary potential, with integration across fields, databases, and agencies, and an overall cultural shift

Note: Additional information on both technical and social opportunities/ challenges is included in the executive summary below as a complement to the NSF Workshop Results summarized above.

Photo by Lauren Cryan, Carnegie Institution for Science

NSF EarthCube Early Career Workshop Executive Summary

Workshop Highlights

At the beginning of the workshop, the participants were provided with an overview of EarthCube efforts to date. Also provided was an illustrative "use case" of the "brute force" connection of data across fields and disciplines (focusing on the labor-intensive process of tracing the formation of Mercury across geologic eras). Among insights and comments that emerged early on in the workshop:

- A focus on incentives to share data and credit the source, including the value of DOIs (digital object identifiers) for data, impact assessments, and funding for data management; appreciation of the factors that would lead people to be "data hoarders," including the time and complexity associated with making data available to others; and a concern that the social systems changes will be more difficult than the technical aspects of EarthCube.

The workshop also featured a presentation of data from a stakeholder survey of 755 geoscientists and computer scientists on cross-disciplinary dynamics and the potential to share data, tools, models, and software. The research pointed to the clear importance to stakeholders for access to data other than their own, but great difficulties in being able to do so. There was variation by field and discipline, with different types of data (field collected, common pool source aggregated, etc.) being a key factor distinguishing fields and disciplines. Connections between the geoscience and computer science communities were revealed as particularly fragile. The most senior scientists reported the least difficulty with data access and the least urgency around interdisciplinary research – a finding with strong implications for the early career scientists since these are their mentors and evaluators. The participants were also invited to use a "career anchors" tool to map their career trajectories and identify the potential for EarthCube relative to their careers. Among the observations that emerged from these two program elements were:

- A deep underlying concern with university promotion and tenure policies, which tend to be conservative with respect to cross-disciplinary scholarship and investments in data beyond the minimum needed to support immediate research objectives.

During the workshop, groups of people in the same fields and disciplines were formed to identify their "hopes" for EarthCube, which included the following selected points identified by the groups (the full detail is included in the report):

- One-stop shopping; improved access to data; ease of sharing data, with standardization; better funding of data storage options; ease in citing data; a "closed circle" from data production, use, review, and publication; aligned ontologies; able to "keep up" with "big data;" minimizing time collating data and maximizing time doing science; "hindcast" and predictive modeling capabilities; international access; access to the general public; achieving the multi-disciplinary potential; integration across fields, databases, and agencies; and a cultural shift in the field.

The participants also identified their "fears" for EarthCube, which included the following selected points identified by the groups (the full detail is in the report):

- Duplication of efforts across directorates and disciplines; disconnect between data and science; data graveyard – useless collection of data; misuse/misinterpretation of data; funding goes to data, not new research; no one in our community wants to take the lead; no incentive structure for publishing data; not enough sustained funding – e.g. support data entry, curation, and storage; creating separation/class stratification between data generators and users; error propagation through datasets; loss of momentum; underutilization; just another hoop to jump through; don't lose the ability to do small projects; suppress novel data collection; intellectual property "violations;" no willingness to collaborate; too rigid or not rigid enough; vulnerability to Cyber-attacks and malicious data use; and lack of differentiation between model-generated and physical data.

One fear, which is that investments in EarthCube would take away funds from investments in research was directly addressed – increases in the level of NSF funding in any given area have generally been to support infrastructure improvements (such as EarthCube), not expansions in traditional research funding.

The design of the workshop was highly interactive – in order to maximize inputs from the participants and to ensure an engaging experience. It was anticipated that cross-disciplinary connections might be made among the participants and that was indeed the case. Part way through the workshop, it was suggested that this community could be a test bed on the sharing of data – so a web-based registration system was set up with the following result:

- Over 48 data sets were identified and described in detail by 13 Workshop participants as data that can be shared now – covering biogeochemistry, biological oceanography, biology/microbiology, chemistry/geochemistry/ chemical oceanography, ecology, education, geophysics, hydrology, informatics/data management, oceanography, physical oceanography, and space sciences.

This is an example of people "voting with their feet" on the sharing of data in the spirit of EarthCube and it an unanticipated, but important workshop outcome.

Key insights from the morning leadership panel included:

- The leadership role of professional societies; the importance of "abductive" reasoning – in addition to inductive and deductive – as an interactive engagement with the data; the emergence of "big science" in geoscience, with the accompanying importance of collaboration; the emergence of a cross-disciplinary institutes at Woods Hole – initially resisted and now functioning well in conjunction with traditional fields and disciplines; the importance and impact of peer-reviewed science, as well as the challenges for scholarly journals looking ahead; the role of journals as repositories for data submitted along with articles for publication

The presentation on governance provided an update on the recommendations for the governance of EarthCube and the signal that it is still in formation. Participants were invited to help shape this process – both at this workshop, on-line, and at upcoming governance events. Key insights from the education panel included:

- The gap between student learning based on student-collected small data sets and professionally-collected large data sets – with the importance of exposing students to professionally-collected large data sets; the value of being directly engaged in field data collection at an early point in your career; the importance of story-telling with data and research findings; the value of staying curious.

Towards a Shared Vision of Success

A short-term (5-7 year) success vision for EarthCube was generated in small groups that were interdisciplinary in composition. Note that the workshop featured attention to the social as well as the technical aspects of EarthCube and that was reflected in the group brainstorming. A summary of the various group recommendations is below, organized into three broad categories (note that these items were constructed from across all the brainstorming list and are offered as a summary – for the exact wording, see the original lists included later in this report):

Access/Uploading:

- Google earth style interface
- Accessible data submission interface
- Standardized meta data on data type, data context, data provenance, etc. for field scientists (with and without internet access)
- Data security
- Public accessibility; empower non-specialists

Utilization/Operations:

- Community mechanisms to build tools
- Large data manipulation, visualization, and animation
- Searchable access by space, time, and context
- Pull up data and conduct analysis with voice commands
- Open source workflow management for data processing and user-contributed algorithms in order to facilitate reproducible research
- Cross-system comparisons; ontology crosswalks for different vocabs in different disciplines
- Easy integration of analytic tools (R, Matlab, etc.)
- NSF support for data management

Output/Impact:

- Mechanisms to provide credit for work done (data, models, software, etc.); ease of citations; quantify impact
- Promote new connections between data producers and data consumers
- Interactive publications from text to data
- Recommendations system (like Amazon) for data, literature, etc.; Flickr for data (collaborative tagging)
- Educational tutorials for key geoscience topics (plate tectonics, ice ages, population history, etc.)
- Gaming scenarios for planet management
- EarthCube app store; ecosystem of apps

A longer-term (10-15) success vision was also sought and the following were among the items identified (some of which could be in the above list):

- First-year grad student can download, manipulate, and model data
- Incentives for release of legacy data, with migration, compilation, and streamlining of access to legacy data
- Data access granularity: confidential data, national security data, and pre-publication data all private until appropriate for release
- Suggestions of additional data to consider
- Full circle: Data includes citations; as data is used in more publications, data is ranked higher
- Peer review of data
- 4D version of Google Maps with "Geosearch" feature
- "open notebook" science
- "facebook" for science data and knowledge
- Bots steaming data
- Workflows for different types/levels of data use and analysis (K-12 to high performance computing)

Next Step Action Items

Next steps following the workshop (also listed at the conclusion of this report) include these relatively quick options:

❑ Sign up to the early career, education and governance EarthCube groups at:
http://earthcube.ning.com/group/early-career

http://earthcube.ning.com/group/education-and-workforce-development

http://earthcube.ning.com/group/governance

(remember to follow the group and also sign up for the mailing list).

❑ Indicate your interest in participating in an upcoming NSF EarthCube domain workshop – they are all listed here: http://earthcube.ning.com/page/earthcube-domain-workshops

❑ For participants, complete the post-conference survey at:
http://earthcube.EarthCube-NG-POST-Consent.sgizmo.com/s3/

Note that this survey is part of ongoing data collection with this community in order to follow the participants (and other invitees to the workshop) as a cohort in connection with EarthCube and the future of the scholarship in the geo and computer sciences.

❑ Plan to attend the EarthCube relevant sessions at AGU in December:
http://fallmeeting.agu.org/2012/scientific-program/

(sessions IN21A, U31A, IN54B, and IN23E or search for "roadmap" in the program).

❑ Post pointers to data you are willing to share at: https://docs.google.com/spreadsheet/viewform?pli=1&formkey=dFNWLWRDd19hSGptLWlHMjZXdXZaaUE6MQ#gid=0

  ○ Contact: Steve Diggs

- ❑ Check for a listing of active NSF solicitations and other funding mechanisms that might be relevant – notices will be sent out when posted
  - Contact: Barbara Ransom
- ❑ Participate in the upcoming Governance webinar
  - Contact: Lee Alison
- ❑ Consider mentioning EarthCube in NSF proposals, but check first on how best to do so
  - Contact: Barbara Ransom
- ❑ Review EarthCube activity to date – what EarthCube efforts have already been funded – 4 community groups and 7 concept awards (web services, modeling, legacy data, cross-domain interoperability, geodata)
  - Contact: Barbara Ransom

More intensive next steps include:

- ❑ Consider organizing an interdisciplinary end user workshop around a theme, as one of the approximately 25 "domain" workshops being supported by the NSF
  - Contact: Barbara Ransom
- ❑ Have another gathering of this group next summer
  - After 1.5 days of dreaming big we should gather as a group and present to ourselves – a self-managed session in which we education each other on the data we work with
  - I would like to take the advice of accomplished people and not organize by disciplines and instead do so by themes – water, carbon, etc. – people can propose themes and then reconvene in small groups and be able to tell success stories
  - Contact: Any member of the leadership team
- ❑ Set up town hall meetings in professional societies to educate and engage people with EarthCube
  - There will be a town hall at AGU – and search for "roadmap" to find three sessions at AGU this year
  - Contact: Barbara Ransom
- ❑ Explore more fully the idea of a professional association around earth science data
  - Either a new association or working under the auspices of an existing one
    - The information science schools have made the choice to not form another society but instead to be informatics divisions within professional societies
    - A similar strategy could apply and journals can be set up under this umbrella
    - It does compete with regular science on the program, but multiple division affiliations are possible
    - International groups also exist, but they tend to be older
    - An independent group could be more flexible and dynamic
  - Contact: Bob Hazen
- ❑ So many of us are itching to do something – a workshop that is a code-a-thon with data

We would need support engineers for a code-a-thon since most of the folks here are in analytics

EXECUTIVE SUMMARY: EARTHCUBE WORKSHOP RESULTS

(Michael Gurnis, Caltech: Oct. 29-30, 2012)

Earth Cube Workshop Title: EarthCube End-User Domain Workshop for EarthScope

Introduction: Supported by the NSF, EarthScope is an ambitious, multifaceted program to investigate the structure, dynamics, and history of the North American continent. EarthScope has deployed a major earth observatory (with geodetic, seismic, and San Andreas fault sampling at depth through drilling) while interpreting and integrating the emerging data. The level of organization and strategic planning within the EarthScope community is high, for example with the community completing a science plan (2010-2020, "Unlocking the Secrets of the North American Continent")-- http://www.earthscope.org/ESSP several years ago as well as preparing a "Preliminary Strategic Plan for EarthScope Cyberinfrastructure" in May of this year (http://www.earthscope.org/es_doc/highlights/ES_CyberinfratructureStrategicPlan_2012.pdf).

We met October 29 and 30, 2012 on the ASU Campus in a workshop organized by the EarthScope Cyberinfrastructure Subcommittee. Attended by 54 participants (reduced by Hurricane Sandy) composed of 25 faculty, 11 post docs and graduate students, and 18 scientists or cyberinfrastructure professionals) that came from a cross section of the community (seismology, geodesy, geodynamics, geology, geochemistry, and information technology/computational science). Before the workshop, we surveyed the community and workshop participants with a variety of questions that spanned science goals, existing cyber tools, roadblocks and needs for new cyberinfrastructure. We received 35 responses to the survey and these results were used for the list given below for the new CI needed for EarthScope science. The science issues and challenges came from the science plan as survey and workshop participants did not make substantial changes to these goals. The excellent presentations fueled wide ranging informal and breakout discussions which led to a number of consensus points discussed below. The final workshop Agenda, slides of the workshop talks, videos of many of the presentations, and some of the posters presented will be posted soon to the EarthScope web site

SCIENCE ISSUES AND CHALLENGES

1. Important science drivers and challenges:

- What is the present-day Active deformation of the North American continent and how is this deformation related to the seismic activity, the growth and activity of faults, and volcanism?
- What is the structure of the North American continental crust and underlying lithosphere and how is the structure related to the present day seismic and volcanic activity and over longer geological times to the assembly of the continent and the record of rifting, collision and maintain building over the entire continent?
- What is the structure of the upper mantle beneath North America and selected regions along the core mantle boundary and how is the structure related to surface geological processes and mantle convection?
- What is the rupture that unfolds during moderate to large earthquakes and how is that rupture related to the state of stress within the crust, the dynamics of earthquakes, rheology of crustal rocks and the presence of fluids within the crust?
- How does the movement of aqueous and magmatic fluids influence the pore pressure, temperature, composition, and rheology of the crust and mantle? How does fluid influence lithospheric deformation and mantle flow?

- Can the EarthScope facilities be used to map water (groundwater, atmospheric water, soil moisture, snowpack, glaciers, and vegetation water content) in time and space in the western United States and Alaska with a resolution that complements other meteorological measurements?

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

- Considerable difficulties exist in finding and accessing data that already exists, including within established databases developed outside of ones immediate discipline. Many of the problems in accessing existing data sets are associated with the enormous heterogeneity of data of interest to the EarthScope community and the standards and formats by which it is stored (spatial vs. temporal, map, volumetric data vs. point data and data from different disciplines -- geophysical, geological, geochemical, meteorological) by which it is stored and accessed through a variety of formats. Consequently, there are no standard interfaces between the numerous data systems. Even for data access, multiple formats lead to substantial hurdles associated with format translation. Even within an existing discipline or a data system, formats and protocols change with time as needs and technology changes. Capturing what has been done to data (including the provenance and all of the complex steps that occur in the generation of higher level data products) can be hard to determine during data discovery and access. There is no universal model definition/dissemination format that has been adopted by ALL earth imaging communities, including the many seismic subdisciplines, but also including electrical properties, density, other properties.

- Because of the enormous breadth of EarthScope science, there is a need to access older (potentially esoteric) datasets that are analog (e.g., maps, geologic paper records, model slices published in paper). There is enormous effort and uncertainty associated with the reverse engineering of ‘raw data values’ from published figures. Data needs to be available independent of publication but hold a publication accountable for its content. Retrieval of data from gray literature and government agencies without a well-developed cyberinfrastructure remains difficult.

- Data integration, a major component of EarthScope science, poses more substantial challenges than eluded to above for data access because of the need to bring a few to many datasets together that individually have unique and complex formats. Common reference frame especially for the spatial integration and visualization of diverse data sets are often lacking and are not necessarily known for those outside of the immediate discipline.

- EarthScope investigators need to bring data in from outside of their immediate area of specialization and the ability to judge and assess errors, uncertainty, reproducibility and consistency associated with raw data and data products at all levels often is entirely unknown. For those outside of discipline, one does not know how the data quality– that is, do specialists rate the data highly, or are there flaws in the data? How do other investigators rate the data, do they find the data useful? There is a need to evaluate consistency/accuracy of existing data? Redundancy needs to be reduced, for example when multiple datasets are aggregated for a model, the workflow/scripts should be made available should be made available so that other investigators can attempt to reproduce and build upon the result. Can data uncertainty be propagate (for example during data integration and generation of higher level products) and can those quantities and concepts be visualized.

- There are considerable problems with the scaling of existing algorithms for big data and the shipment and movement of datasets between datasets and processing locations. Investigators need access to HPC platforms beyond their immediate research groups and universities and investigators cited concerns with the long queues that exist with current facilities and the administrative effort associated with gaining resource allocations.

- EarthScope investigators cited concern with the access to software engineers and IT specialists with appropriate skill sets to allow them to solve data access, data integration and knowledge product generation. There was also concern with how IT specialists can partner with domain scientists. How can one find the overlap of interesting topics between domain and IT? There was concern on how to move beyond the prototype (which may exist at a research level in computer science or an IT field) and the development of the technology and methods so that it can be used for production

- Despite the wide available of open data and open source software, some concerns with proprietary data and software remain. In particular, there may be needs for more incentives for data contribution. Specifically, data producers remain concerned with "getting scooped" after making their data available (open data) and special protections might be needed for early career scientists.

TECHNICAL INFORMATION/ISSUES/CHALLENGES

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

- Deployment of simple web services across several domains with a sophisticated brokering system(s).

- Workflow with standard interfaces to the underlying components. Initial prototype and then produce script

- Report, propagate, visualization of uncertainties . Tools for model validation and assessment (misfits) of available models. "misfit comparison between these models is …"

- Community driven evaluation of data

- Enhanced access to underlying models and data in publication. Possibly require submission of datasets to repository as part of publication. (include workflow/scripts associated with data) "carrots and sticks" to encourage sharing easy uploading. publishers/checkbook requires uploading

- An enhanced, but simplified, open source GoCAD for the 3D and 4D spatial integration of geological volumes, points, lines, and surfaces. An Earth model construction environments that allows information to be correlated and plotted will extracting and calculating additional quantities. Visualization tools for 3D and 4D datasets. Comparing different 3D models quantitatively. Common framework to share and compare models. This should allow one to overlay different data/models geographically integrate petrological models and data seismic model library

- A web-accessible plate tectonic reconstruction system (paleo-GIS) that allows the earth to be typified by a hierarchy in the scales of deformation, from global rigid plate motions, to regional deformation with local faulting along with paleogeographic reconstructions is required.

- Data analysis software packages that are well connected to the data center(s). An important component would be standard data mining and pattern matching routines

- Extend usefulness of the Compuational Infrastructure for Geodynamics (CIG) for EarthScope-science. This could involve the development of an environment for inverse problems.

- Much greater collaboration between cyberinfrastructure development in EarthScope and earthquake early warning activities

- Deployment of robust hosting services with a distributed architecture. The system would need a rich privacy and permission control over content to facilitate sharing or restricting by users and user groups (sophisticated content management system). Provide repository for data and policy designed to encourage/coerce sharing that data.

- Funding for domain experts to collaborate with computer scientists. Allocated funding which provides support for researchers who need programmers and a pool of "certified" EarthCube programmers. In general, I think development of new tools should be driven by the users. The most effective approach is to have programmers clean up tools initially developed by individual researchers to make them user-friendly. There are too many examples of tools initiated by programmers that end up being of limited use to the community because they were ill conceived from the start. "My main limitation is time to learn to use tools that are already out there." 2) A virtual institute for community software in seismic and MT (and related fields)

- Training and documentation for software and data centers. Involve users in the documentation process. Wiki or living document, user forums, and social networks are ideal mediums for communicating with the user base.

- Standards APIs for querying 3-D seismic velocity models and flexible data structures for their representation that facilitate large models (10^8-10^10 grid points) and fast querying. Web portals for simple querying of 3-D seismic velocity models are needed to provide earthquake engineers with the parameters they need for simple ground-motion prediction models. Techniques for representation of epistemic uncertainty and small scale features in 3-D seismic velocity models and 3-D fault models.

- We need to be able to access as much geoscience data as possible through the "cloud" and in the field. This requires vertical integration of datasets, where information is sorted/queried by location.

- Easier connection between tomographic models and wave propagation codes would be helpful. If IRIS EMC is the standard, then we need to adapt our codes to readily read in these models. 2. "Push-button" assessment of tomographic models, based on running a suite of independent earthquake simulations, then calculating various misfit measures.

- Powerful client side applications to access diverse datasets in user-specific ways to conduct analysis and visualization (MATLAB would be a good substrate for this) Use-case oriented term and units translations between diverse datasets as is applicable to specific studies, represented as ontologies. Uniform, open, REST-oriented web services with domain specific terms and data, but tools to promote translation to other areas of study, as opposed to being simply homogenized to the lowest common denominator.

- 3D seismic forward modeling with setup tailored by user - geographically indexed database with results from all geosciences and links to journal papers in ISI - cross disciplinary data access (at stages between raw data and published results)

- Convenient and flexible interface for students to browse and manipulate seismic data. 3D seismic wave simulations at continental scale and periods <20 s.

- On-demand processing environment to produce higher level products from raw SAR data (interferograms, InSAR time series). 2. Archive of higher level InSAR products

- A user-governed model based on a simple API with data discovery and visualization capabilities in the 4D space, which would allow users to submit their own data, models and workflows. All version controlled and semantically enabled. 2. Scripts with a clearly defined syntax that could immediately make any program part of the system: one could select of subset of information, "click" and execute a workflow. In this setting, codes could be run by a user on selected data directly on the relevant remote servers, through the API. 3. Submission of research products to EarthCube could be part of the NSF data policy, while sharing controls could be set by users on their personal content. Contributions should be optionally peer-reviewed, citable, and author-tagged through an underlying social network.

- Metadata descriptors that allow facile query of databases database exploration tools fast networks for transferring large quantities of data

Summary

EarthScope community is ready to tackle the technical challenges we identified in this workshop and transform its scientific practice and development of geoscience knowledge. The EarthScope community is extremely diverse while simultaneously being coherent through its focus on the North American continent and a series of bold grand-challenge questions that we have previously refined and articulated. Moreover, the community has a wide range of existing CI facilities and IT-agile academic partners that are poised for the next step of geoscience-wide data and knowledge integratation. We plan to respond to EarthCube requests for proposals.

EXECUTIVE SUMMARY: EARTHCUBE EDUCATION WORKSHOP

(K. Kastens and R. Krumhansl, Education Development Center, Inc.,
and Cheryl Peach, Scripps Institution of Oceanography)

(report drafted: March 15, 2013; revised May 7, 2013)

Earth Cube Workshop Title: EarthCube Education End-User Workshop

Introduction:

Forty-six geoscientists, geoscience educators, data providers, employers, technologists, and curriculum developers met on March 4–5, 2013, at Scripps Institution of Oceanography to advise EarthCube's leaders and builders on the needs of end-users who will use EarthCube for education. The learners targeted by our recommendations include traditional undergraduates, but also other motivated adult learners, especially scientists from other disciplines collaborating with geoscientists on interdisciplinary problems. The goals of the workshop were:

- to build EarthCube in such a way as to bring the power of learning through geoscience data and models within reach of novices
- to use EarthCube to educate future geoscientists, who will be unprecedentedly facile with data and models, and "native speakers" of interdisciplinary systems

GEOSCIENCE EDUCATION ISSUES AND CHALLENGES

1. **Important geoscience education drivers in the 21st century:** (list 3 to 6).
   - Few of the big claims of geoscience (e.g. plate tectonics, global climate change, age of the Earth) can be explored in traditional student laboratory activities in the way that physics students can experiment with forces and accelerations or biology students can experiment with growing plants. Thus, the availability of professional-caliber datasets on the Internet has been transformative, insofar as it has allowed geoscience students to engage, in many cases for the first time, with the data that form the evidentiary basis of the concepts that they are studying. Geoscience education is out ahead of the other sciences in its use of large professionally-collected datasets for undergraduate education, and thus is having to pioneer new pedagogy around teaching and learning with data and models.
   - We live in a data-infused society. In today's workforce, data isn't only for scientists. In an ever-increasing percentage of professions, from nurse to car mechanic to teacher, adults are expected to be able to make use of data in the daily demands of their work. In our current education system, science and math are the places where students encounter data, and so teaching basic data using skills ("data literacy") in these classes has become a basic workforce training imperative for all students.
   - Beyond the level of basic data literacy comes a degree of mastery that the workshop participants referred to as "data-savviness." The workshop wove an inspiring vision of the attributes of a data-savvy college graduate, skilled at using data and models to answer difficult questions and solve hard problems, facile with systems thinking and interdisciplinary problem solving, and able to make inferences about process and causality from Earth observations.
   - Our society faces many difficult decisions in the 21st century. The workshop conveners and participants think that better decisions would be made if a larger fraction of the populace used evidence, evidence grounded in data, in making decisions in their personal and professional lives. Undergraduate education is a prime time to establish the habit of mind of using data as an input to answering questions or solving problems. For the suite of decisions that revolve around approaching limits to growth on a finite planet (energy, water and mineral resource limits; environmental degradation; climate change), the relevant data will be of the sort served by

EarthCube.

- As computational models become a much more important part of the geoscientists' toolkit, geoscience education is endeavoring to convey this trend and prepare students to be a part of it. This is proving difficult, in part because students bring forward with them from K-12 the expectation that models are for demonstrating or explaining that which is already known rather than hypotheses to be tested by comparison with data. The epistemology of how scientists actually go about creating new knowledge using external runnable models is not widely understood by teachers or by the public.
- Pre-college education is on the cusp of change, driven by the advent of the Next Generation Science Standards (NGSS). The NGSS foreground "science and engineering practices," including "analyze and interpret data" and "develop and use models." If and when the NGSS are fully and widely implemented, students will arrive at college with much more knowledge of the Earth, of data and of models—and with an expectation that science education should involve activities in which students construct meaning through active exploration. In the meantime, the prominence of the practices the NGSS is spurring a flurry of education research on the practices, including the data-using and modeling practices.

2. **Current challenges to high-impact, interdisciplinary geoscience education using data and models:** (list 3 to 6).

- Tools and interfaces designed for geoscience experts pose a significant barrier to use by students and other novices, especially when working on interdisciplinary issues.
- Many students enter college with minimal knowledge or skills around data, models, or the Earth.
- For some topics and audiences, there is a shortage of high quality instructional materials (especially around models, but also around data), that involve students in active inquiry rather than cookbook direction-following.
- Many instructors lack pedagogical content knowledge (knowledge of how to teach a body of content rather than knowledge of the content itself) around teaching with data and models, and lack a community of practice within which to develop this knowledge.
- Relevant cognitive/learning science is sparse and insufficiently incorporated into instructional design.

TECHNICAL INFORMATION/ISSUES/CHALLENGES

1. **Desired tools, databases, etc., needed for geoscience education:**

- The undergraduate geoscience education community makes uses of a very wide variety of geoscience data types. To see the range and depth of data in current use in geoscience education, please browse the following collections: *Using Data in the Classroom: Data Sources and Tools: (*http://serc.carleton.edu/usingdata/resources.html) and *Earth Exploration Toolbook Chapters:* (http://serc.carleton.edu/eet/chapters.html).
- Cyberinfrasture desired by the education workshop (see further detail in full report):
    - Data germane to society's pressing problems
    - Field data
    - Ability to ingest and display student collected data
    - Tiered approach to data quality (allowing quality student data to be added, while keeping out inadequate data)
    - Near real-time data, and also historical archival data
    - Local informants' eye witness accounts
    - Novice-friendly interface options, scaling gradually up to the full professional interface
    - Support for data exploration and "making 'failure' cheap"
    - Collaboration tools
    - Supports for understanding uncertainty in data and model output

- Comprehensive and comprehensible metadata
- Simple and well-documented versions of geoscience models
- Support for student building of models

- Pedagogical & social infrastructure desired by the education workshop includes:
  - Support for citizen science
  - Mentoring for both teachers and students
  - Assessment techniques for student mastery of data and modeling practices.
  - Tutorials and training sessions
  - Venues in which to share and build a community of practice
  - Support for diverse populations, including learners with disabilities, urban youth with limited access to nature, and adult professionals crossing fields.
  - Support for entrepreneurial enterprizes

COMMUNITY NEXT STEPS

1. **List of what your community needs to do next to move forward and how it can use EarthCube to achieve those goals:**
   - Work with cyberinfrastructure designers to ensure that EarthCube's tools and interfaces have options to present themselves in ways that make sense to, and communicate effectively with, learners and novice users (see details in full report).
   - Work with in-service K-12 teachers and pre-service teacher educators to ensure that the EarthCube-relevant portions of the Next Generation Science Standards are implemented effectively: the Earth & Space Science Disciplinary Core Ideas, Practice 2 ("Developing & Using Models") and Practice 4 ("Analyzing and Interpreting Data".) EarthCube-based lesson plans, and EarthCube's social infrastructure (webinars, workshops, social media, etc.) will be useful in this effort.
   - Lay out a learning science research agenda covering key questions in how humans learn with data and models. Collaborate with cognitive scientists, learning scientists, and education researchers to implement the plan, using EarthCube data as a testbed for some studies. Disseminate the findings among curriculum developers and faculty.
   - Continue to develop and test teaching and learning materials that involve students in making meaning from data and models, using instructional sequences that go beyond cookbook step-by-step. EarthCube can help by providing virtual and face-to-face venues where such instructional materials can be shared, vetted, and improved.

EXECUTIVE SUMMARY: WORKSHOP RESULTS

(Liping Di, George Mason University)

Earth Cube Workshop Title: Engaging the Atmospheric Cloud/Aerosol/Composition Community

Introduction: Scientists working on the atmospheric cloud/aerosol/composition (ACAC) domain typically develop theories, models, and predictions on the state and dynamics of the atmosphere and its constituents by acquiring, processing, analyzing, integrating, assimilating, and modeling with data from diverse, multi-disciplinary sources, both in-situ and through remote sensing methods. The volumes of the data used in the research can be small or very large ("big data") and the data could be from live sensors, archived at the big data centers, or at the hand of individual scientists. Such diversity on the data poses great challenges to ACAC scientists on their research and education activities. Therefore, common cyber-based infrastructures, such as EarthCube, for handling and managing diverse data and facilitating information extraction and knowledge discovery from the data are urgently needed. The purpose of the workshop is to gather community inputs on current challenges and requirements to the EarthCube.

A total of 67 scientists from the ACAC community participated in the workshop. Among all participants, 60% of them are from universities and 40% from government agencies, industry, and non-profit research centers. All of the participants are affiliated with domestic organizations of the U.S. The main outcomes of the EarthCube workshop discussions are summarized below.

SCIENCE ISSUES AND CHALLENGES

1. **Important science drivers and challenges:** Participants identified several high-priority science questions that will be the focus of interdisciplinary efforts during the next 5-15 years.
   - What are the sources and the removal mechanisms of chemical species in the atmosphere?
   - A lot of work has focused on improving Ozone but other species, for example, Methane and NO_x, have been neglected in the process. The entire atmosphere system is sensitive to changes in NO_x and needs to be considered. What are the effects of industrial NO_x on atmosphere composition?
   - What are the exact roles of the clouds in the cloud systems and in the entire earth system? Several outstanding cloud-related deficiencies in the climate modeling are well documented by the research community and need to be addressed in the next 5-15 years, including the double ITCZ problem, poor MJO, too short and too regular ENSO periodicity, diurnal cycle and frequency of precipitation, inconsistent representation of radiation and clouds.
   - How do clouds affect the cloud feedback on climate sensitivity?
   - What is the role of clouds on biosphere or ecosystems or vice versa?
   - What is the spatial, temporal, size distribution and composition distribution of aerosol particles in the atmosphere and the aerosol particle emissions globally?
   - What are the exact roles of aerosols in the cloud and climate?
   - What is the impact of aerosol on severe marine storms?
   - What are the changes to Cloud Condensation Nuclei (CCN) with changes in aerosol loading?

2. **Current challenges to high-impact, interdisciplinary science:** The participants, through mentally exercising inter-disciplinary research procedures for solving the above mentioned science challenges, came out a set of consistent barriers and challenges that prevent the above-mentioned science challenges from being solved easily.

- The inter-disciplinary research requires the use of diverse data from diverse sources. Significant challenges still remain for scientists to discover, access, integrate, and use those data in their research.
- Long-term Earth observation through remote and in-situ sensors is one of the major data sources for the inter-disciplinary ACAC research. However, the continuity of satellite & sensor is an issue since the current fleet is aging and not sure what the future holds, including the transition from research-based satellites (NASA) to more operational (NOAA) platforms. Another issue is how to obtain and use data from satellites of other countries. Currently there are multinational efforts (e.g., Europe, US, and China) on satellite measures of ACAC. Enabling the interoperability of data between those efforts and the U.S. efforts is a challenge.
- More data products and/or higher data resolution mean a lot of more data needs to be transferred. This introduces the bandwidth issue both for satellite downlinks and at the user-end (internet). Increased onboard computing power (compression, data sampling) may address this. Combining operational and research platforms is a challenge. The research community is competing with an exponentially increasing data hungry society (Netflix, online streaming, telecommunications, remote working, etc.).Who pays the bills for the mounting cost of the network backbone support?
- Comparing to satellite remote sensing data, in-situ data are less accessible (but high value content) data. Most often it is only accessible by personal requests, making it very hard to develop a harmonized dataset for regional and global models. The model data are by far the least accessible (but also high value content) data. The un-accessibility of model data makes the model diagnoses and inter-comparison difficult.
- Inadequate metadata on data quality and provenance makes scientific use of data from other sources difficult since scientists have hard time to understand if the obtained data are useful. No standard is available yet as to when data is useful (some properties, e.g., ice crystal number are orders of magnitude off, so perhaps that may not be useful)
- The inter-disciplinary research often requires integration of data across sensors and platforms. However, not much has been accomplished in this area yet. The major issue is co-location of sensors with each other and models with sensors. Recent progress on sensor web technology may relieve this issue a bit.
- Many needed global datasets are not yet available or with bad quality. An example of such datasets is the cloud hydrometeors (crystal number, droplet number and cloud phase) for statistical evaluation of models. Aerosol cloud particle precursors climatologies are lacking (CCN, IN) but that is becoming better. Vertical velocities are critical but nonexistent. Cloud lifecycle datasets are nonexistent. The lack of datasets impedes our ability to evaluate conceptual models of cloud development and aerosol-cloud interactions.
- Significant insider knowledge on the data and IT skills are needed for full utilizations of these data resources. No adequate tools and services are available for readily integrating data from multiple sources and across disciplines.
- Modeling is the major method in the inter-disciplinary ACAC research. Data and products obtained from Earth observation sensors are extremely useful in the model initialization, verification, and validation, and as model constraints. However, it is a challenge to make sensor observation data easily consumed by models. Common methodology, framework, and tools for easy sensor-model coupling and integration are needed.
- Lack of community standards (or too many standards) on data format, file types and metadata make the interoperability and sharing of interdisciplinary datasets difficult.
- Many current researchers have not been trained to collect and report data in an interoperable way so that it can be reused by others. There is also lack of formal mechanism for rewarding scientists who share their data, algorithms, and services.

TECHNICAL INFORMATION/ISSUES/CHALLENGES

Data Access Challenges

- Different types of users need different types of support (some, for example in developing countries, just want to import data into Excel)
- Cross-community access is the biggest challenge (within an domain community, it is generally understood how to work with data formats and tools)
- Need to understand the data characteristics (quality, provenance)
- Enable scientists to find relevant, reliable data regardless where the data are archived and obtain the data in the form specified by the scientists
  1) Search by location, date, topics, etc.
  2) On-line services for providing automated data customization
  3) Globally available data; data in other country's agencies
  4) Able to integrate data from different platforms and repositories
  5) System interoperability (inter-agency and international sharing)
  6) Better metadata and standards for data understanding and usage
- Can EarthCube provide a better search engine tailored for communities?

Non-domain Understanding

- Challenge is in having users understand the uncertainty and errors associated with data.
- Documentation is critical. Needs to be understood by others outside of the immediate discipline that created the data.
- Data processes change over time
- Need education/training about data (perhaps a service EarthCube could provide?)
- Can EarthCube fund training for cross-disciplinary data management and informatics?

Supporting small datasets in EarthCube

- Many groups, (e.g., research labs) have small, individually maintained datasets and do not have a large infrastructure to support them in the publication and management of them
- EarthCube should support these small datasets
- One example, might be an EarthCube sponsored cloud data management and publication service to simplify the process for smaller groups

Supporting Data Management

- Challenge is in funding data management - for example, many research groups don't have the funding or time to do metadata creation.
- Interest in an EarthCube supported cloud service or tools
- Employ people within EarthCube who have library science and similar skills to help organize and provide access to data

Data Quality and Standards

- Need a set of EarthCube recommended standards and best practices to facilitate the interoperability and sharing.
- Bad data needs to be flagged
- Need a rating system to help determine and convey what is the quality and type of published data.
- User need to understand when they're using data at their own risk or when it is peer reviewed
- Need mechanisms to catch uncertainties and errors in data before and after they are published
- Unique dataset IDs are created to link datasets to publications and datasets to each other, Suggestion to only provide a dataset an ID if it is peer reviewed and determined to be acceptable.
- Should provide supporting documentation that describes how dataset was derived (algorithms, software used, etc.). And need to track data processes as they change over time.

Supporting modeling and integrated analysis

- On-line data integration and analysis services
- Tools and services to manage, archive, and disseminate model outputs for facilitating modeling comparison
- Sensor-model coupling for facilitating model verification and validation with observation data
- Sensor web and models as services

Merging Existing and New Infrastructures

- Need to transition existing systems to EarthCube, not requiring an overhaul of existing systems
- Need translators, converters, adaptors
- Strive for common standards and practices where most effective

Assessment of current tools and cyberinfrastructure capabilities/best practices

- Provide testbed of developing cyberinfrastructure components, tools, systems, and etc.
- Provide tools to help professors incorporate new datasets and tools into classes.
- Create an EarthCube working group to be involved in assessing value and usability of tools and improve teaching with actual data.
- Focus on the establishment and enforcement of metadata standards.
- Enforce the use of common accepted metadata language.
- Focus more on the development of extended metadata and less on data formats.
- Support the development of converts, translators and readers from one data format to metadata and back to data in a different format.
- Define the languages to read metadata.
- Support interoperability for hardware and software components.
- Improve understanding of users and their needs: students, scientists, public, government agents, university professors, etc.
- Maintain a standard digital object identifier (DOI) system. This is a character string (a "digital identifier") used to uniquely identify an object such as an electronic document.
- Improve access methods. Focus on the development of tools and schemes that help users to find the data and products that they need and show how to access them.
- Is it possible to build one step for all?
- Provide information and access to data properties and quality.
- Favor interdisciplinary approaches.
- Search for ways and methods for multi-sensor, data and products combination.
- Need for instrument simulation capabilities.
- Allow integration and retrieval of data and products.
- Favor long-term continuity of data and products (the issue of trust and lineage between different data sets).
- In case of data/products gap, offer ways to fill them.
- Provide tools for data/products discovery, analyses and integration.
- Enforce documentation standards.
- Provide visualization tools.

Making scientists easy to contribute their data and sources to EarthCube

- A set of easy-to-use tools for scientists to document and publish their data and cyber-resources (e.g., algorithms, models, and computing facilities) for sharing
- The academic community needs to change the way for valuing the academic achievement. Researchers should get credit for sharing their data and resources.

EXECUTIVE SUMMARY: EARTHCUBE WORKSHOP RESULTS

Convened by A.K. Aufdenkampe, C.J. Duffy, G.E. Tucker
Univ. of Delaware: Jan. 21-23, 2013

Earth Cube Workshop Title:

Engaging the Critical Zone community to bridge long tail science with big data.
Short Title: Critical Zone EarthCube Domain Workshop

Introduction:

Critical Zone (CZ) scientists take as their charge the effort to integrate theory, models and data from the multitude of earth science disciplines collectively studying processes on the Earth's surface - from the atmosphere at the vegetation's canopy to the lower boundary of actively cycling groundwaters. Sixteen CZ disciplines were represented at the workshop with experiences that span the range from Big-Data to Long-Tail science. The national Critical Zone Observatory (CZO) network has begun the process of building a community cyberinfrastructure that the workshop organizing committee feels can serve as **a pilot for the EarthCube endevor** of engaging of a diverse community of Earth System scientists to embrace and co-develop a shared data and modeling system.

The Critical Zone EarthCube Domain Workshop had 103 registered participants, with 68 participating in person, 40 participating virtually, and several on-site visitors. 28 were early career (6 graduate students, 11 post-docs, 11 assistant level faculty). Most participants were self-described as representing more than one of the 16 CZ disciplines:

● Biogeochemistry	30
● Biology / Ecology	15
● Biology / Molecular	3
● Climatology / Meteorology	15
● Data Management / CyberInfrastructure	46
● Engineering / Method Development	8
● Geochemistry/Mineralogy	13
● Geology / Chronology	14
● Geomorphology	15
● Geophysics	8
● GIS / Remote Sensing	31
● Hydrology	46
● Modeling / Computational Science	36
● Outreach / Education Research	7
● Soil Science / Pedology	16
● Water Chemistry	14

Workshop participants self-divided into four breakout groups, developing extensive responses to workshop outcomes over five breakout sessions. We have here distilled those breakout notes, which are available at https://drive.google.com/#folders/0B_VW4kvIBAzQSE91SEdkVzIGYkE.

SCIENCE QUESTIONS, CYBER CHALLENGES AND NEEDS

1. **KEY SCIENCE QUESTIONS:** Participants identified several high-priority science questions that will be the focus of interdisciplinary efforts during the next 5-15 years (compiled from all breakout groups, sessions 1 & 5):

The central scientific challenge of the critical zone science community is to **develop a "grand unifying theory" of the critical zone through a theory-model-data fusion approach.** This concept expands on the classical notion of Hans Jenny's state equation for soil formation -- *S = f(cl,o,r,p,t,….)*, where *S* is for soil, *cl* represents climate, *o* organisms including humans, *r* relief, *p* parent material (lithology), and *t* time -- into a 4D landscape-scale model of coupled physical/ chemical /biological processes that frame the critical zone's evolution, function, and response to change. Developing such a grand unifying theory requires answering three broad questions:

1. How do tectonics, lithology, climate and biology co-determine the evolution of critical zone structure and function?
    - "structure" = 3D arrangement of the remnants of physical and chemical weathering from surface to bedrock, and associated spatial patterns in biological communities. It includes properties such as topography, chemical composition, porosity, permeability, and physical structure (cohesion, fracture density, shear strength, and similar properties), as well as biological communities both above and below the surface.
    - "function" = the processes of transforming and transporting energy and materials. CZ function includes all "ecosystem services", including water routing, storage and filtration; biogeochemical transformations such as nutrient, carbon or greenhouse gas uptake/storage/release; sediment flux; and others.

2. What are the drivers of energy and material fluxes (i.e. water, sediment, carbon, nutrients, solutes, etc) moving through the critical zone?

3. How will critical zone structure, function and evolution respond to human and natural disturbances and over various time and spatial scales?

A second, yet equally important, challenge is whether a unified theory of the Critical Zone can create the necessary knowledge base to evaluate the complex issues of supporting sustainable landscapes. Several specific high priority questions were identified to provide detailed examples of the applications of the broader questions above. These were considered high priority in large part because of their immediate relevance to human and ecological sustainability issues.

- What is the impact of human-induced changes to the nitrogen cycle on the land, air, water, and ecosystem of the critical zone across the scales where science-based management decisions and actions are made (individual land parcels to basin scales)?
- What is the current distribution of soil carbon at global, regional and landscape scales, and given the drivers of these distributions how will soil carbon stocks change in the next 50-100 years?
- What essential biodiversity and other biological variables are most relevant for characterizing the biological processes that co-determine critical zone structure and function? At what scale are these variables best measured?

2. **KEY [CYBER] CHALLENGES:** Several themes emerged as consistent challenges faced within/across the involved disciplines (compiled from all breakout groups, sessions 2 & 5):

General cyber-challenges include:

- CZ data is diverse and much of it is “dark” . There is no one-stop shop for even knowing what is available, let alone accessing it.
- One constraint that limits community access to Essential Terrestrial Variables (ETVs) for watershed modeling is that the data sit on many servers, with multiple (and heterogeneous) formats, very large files, and complex security, making it difficult for scientists or students to use the data. A second challenge is that even if the above problems were fixed, the scale of the data and the tools necessary for data mining, fusion, and visualization are not yet readily available or usable by scientists. The problem of accessing and sharing real-time data collected by CZO scientists is a theme in this challenge
- Modeling, computation, and numerical prediction is carried out in an ad hoc manner with limited cross-domain collaboration (water-bio-rock) and without the benefit of close interaction with cyber scientists and numerical analysts. An outcome is that such results both challenging to obtain and are not easily reproducible.

Specific scientific challenges that require cyber-solutions:

- Understanding diverse scientific worksflows by CZ scientists and applying appropriate tools to promote shared discovery requires a fundamentally new approach to how the scientific process will evolve from experimental data, to interpretation and models, to the creation of knowledge and wisdom.
- Uncertainty and variability are fundamental to all CZ use cases. Across a range of activities -- from field experimentation where sensors are impacted by environmental noise, to issues of communication in wireless sensor networks, to real-time data assimilation in nonlinear spatially distributed models, to data and model analytics, visualization and computational steering -- uncertainty and variability must be addressed. Although these areas are effectively dealt with in individual CZ disciplines, there is not at present a general framework to efficiently deal with this specific challenge.
- Closed technologies such as WSN’s (wireless sensor networks) have evolved as proprietary products that are not yet useful for the Critical Zone problems where low-power, integrated, heterogeneous, co-located systems of research-grade sensors are necessary to resolve multi-state, multi-process discovery within fully coupled bio-geo-chemical hydrological systems. In particular, research-grade, low-power bio- and chemo-sensors are particularly missing in the integrated measurements at CZO’s.

3. **KEY [CYBER] NEEDS** needed for pursuing key science questions with brief elaboration (compiled from all breakout groups, sessions 4 & 5):

Each breakout group independently envisioned a future cyber-infrastructure that might enable seamless 4D visual exploration of the knowledge (data, model outputs and interpolations) of critical zone structure and function, similar to today’s ability to easily explore historical imagery of the earth’s surface using Google Earth. This map-based visualization system would allow a user to zoom above or below the Earth’s surface to view:

- point locations with sensor-based or sample-based time series observations, and direct access to that data

- profiles from soil pits and boreholes with sample-based data, and direct access to that data
- 2D satellite imagery and GIS data coverages from many different agencies and sources, with time sliders to explore historical images and view differences in time
- 2D & 3D images of CZ structure obtained for the subsurface via geophysical approaches or for the surface obtained via LiDAR and other geospatial imaging approaches.
- Depth to groundwater and depth to bedrock
- Modeling results, visualized in 2D, 3D and 4D
- 2D and 3D interactive visualizations of select datasets

A number of immediate needs were identified:

- Much more data in discoverable repositories with full metadata (i.e. too much of CZ data is “dark data”).
- Easy-to-use web application suites for integrated data discovery, access, visualization and publication. This would lower the activation energy to get more CZ scientists to use the existing cyber-infrastructure.
- A power users “toolbox” in an easy-to-install and easy-to-teach cross-platform package to enable cyber-savy CZ scientists and data managers to more easily manage local data, publish their data to repositories and directly access existing data resources using web services APIs.
- Training and support to increase the overall computational and data handling skills of the CZ science community at all levels, from taking the first steps beyond spreadsheets to contributing to open-source scientific software projects.
- Central data catalogs that allow single searches of multiple repositories from many CZ disciplines and domains.

EXECUTIVE SUMMARY: EARTHCUBE WORKSHOP RESULTS

(Jennifer Arrigo, CUAHSI and Ying Fan Reinfelder, Rutgers: January 29-30, 2013)

Earth Cube Workshop Title: Hydrology/Dark Data – Envisioning a Digital Crust

Introduction: This workshop brought together geoscientists to develop a community vision and a path forward to achieve a "Digital Crust" – a three-dimensional digital representation of the composition and structure of the continental crust of North America that would advance our ability to quantitatively describe, model and understand fluid flow in the subsurface, from the critical zone to the deeper crust. While the primary background of participants was hydrology, and those interested in modeling water flow in the critical zone at local to global scales, some participants came from other geosciences and research areas such as geophysics, geothermal energy, stratigraphy, geochemistry, and geodynamics/deep crustal fluid flow. There were 43 on site participants, and 18 virtual participants via WebEx, mostly (74%) from US based institutions. The stakeholder survey indicated that many of the participants were experienced researchers (74% having over 11 years of experience in the field). Main outcomes from the workshop are listed below.

SCIENCE ISSUES AND CHALLENGES

1. **Important science drivers and challenges:** Participants identified several high-priority science questions that will be the focus of interdisciplinary efforts during the next 5-15 years (list 3 to 6).

   - A high priority is understanding the evolution and functioning of the earth's critical zone, defined as the thin near-surface layer of the crust that sustains all terrestrial life. Fluid circulation and thus enabled energy, carbon, nutrient and other geochemical fluxes play a critical role in shaping the evolution of terrestrial biosphere and societies. The structure of shallow groundwater flowpaths, and its exchange with surface waters and the vegetation root-zone determine the seasonal water availability to vegetation and aquatic ecosystems, as well as carbon and nitrogen transformation and transport. There is no information on the material properties below the soil survey depth (~1m), preventing interpretation of field observations and modeling efforts across cm to watershed to regional scales.

   - Another high priority for our science is to advance a synthetic understanding of forcing of groundwater flow over many scales. Currently groundwater assessments are done at discrete scales, and information is not typically transferred between scales (upscaling or downscaling). The digital crust effort could provide a means to evaluate forcing of groundwater over a very wide range of scales (local, regional, continental), and to understand linkages between scales (e.g., effects of changing precipitation patterns and sea level on regional-to-continental groundwater levels, cumulative effects of water withdrawals, effects of regional-scale modifications of land use and surface drainage networks), as well as provide the basis for better incorporating groundwater in earth system models in ways that allow us to evaluate two-way feedbacks between groundwater and climate system on much larger and longer timescales then currently possible.

   - Another fundamental science question that bridges several geosciences disciplines and has extreme relevance for society is understanding the role of fluids in seismicity and tectonics.

How can we quantify the distribution and magnitude of fluxes from the brittle to the ductile regime; can we better under the interaction between the hydrosphere and lithosphere?

- Share different interpretations of available data into geologic structures. Data standards and tools are not currently adequate to allow domain scientists to share interpretations or to quantitatively compare and contrast different interpretations of the various kinds of geologic, geophysical, and mineralogical data used to infer geologic structures.

- Organize the variability, connectivity, averaging, and covariance of disparate physical and chemical properties of the crust within the context of geologic structures. One of the central challenges identified by having workshop participants discuss their knowledge and challenges within their own disciplines is the fact that the earth appears to have many structures depending on the particular properties used to define the structure, but many applications require synthesizing information on multiple properties (e.g., weathering → temperature, mineralogy, water flow and chemistry, etc.; nitrogen dynamics → temperature, water flow, oxygen, carbon, microbial community structure, active microbial biomass and/or metabolism). At the current time, we don't have a good sense of how these various representations of earth's structure compare, or at what scales different properties average, or how important properties co-vary (or don't co-vary). We also need to directly face the fact that all estimates of structure have a high degree of uncertainty.

- Advancing our understanding of paleo-reconstructions of depositional environments. A specific example discussed at the workshop was the Gulf Coast. Building a complex model of the 3D geology of depositional environments over several periods of time (from Mesozoic to present day). Mapping these over time gives a much better understanding of the complex stratigraphy of these depositional environments, allows targeted sampling of geologic features to derive source evolution for tectonic investigations, and can aid societal needs such as energy exploration.

- Another high priority science challenge identified by some participants is to further our understanding what the geologic, geomorphic, and environmental factors that determine the formation of the unique environments – e.g. karst systems. Karst systems exemplify the type of transformative and societal important research the digital crust would enable. Karst covers about 20% of the earth's surface, and are incredible fragile environments - subsidence (natural hazard), water quality, and urban/other planning that needs to understand the impacts of karst geology on water supply, construction, and other issues. Researchers currently studying karst environments often have to create their own datasets from many disparate and regional sources, and these studies often create a wealth of data that are not easily shared, so we do not have a comprehensive picture of global karst research and information.

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).

   - The major theme that emerged is that in all these applications we are dealing with a poorly observable system. The constant challenge is how to represent different interpretations and capture the knowledge of the scientists that created them, and to quantify and deal with uncertainty. In a discussion focused on primarily hydrologically-relevant properties, it was discussed that there could (would) be multiple geovolume interpretations. Sharing knowledge across disciplines would require substantial metadata or context. A central hypothesis of this effort is that a reference geologic framework will be useful to organize attribute data;

simplifications of this geologic framework can be done for different hydrologic or geochemical applications.

- Another challenge in dealing with an poorly observable system is that most data are inferential: for example, geophysical logs can be interpreted to indicate geologic formations but inversion algorithms are subjective. The type of data used to even create a simple 2D or 3D geologic map includes several data types (well logs, seismic reflection data, samples, published and paper references) that must be integrated.
- A central challenge in creating the digital crust is the issue of dark data. Much of the data investigators are using for their research are not digitized. The baseline subsurface information for much of the continent is sparse. Many participants stressed the amount of time and effort it took to track down and assemble suitable data. In addition, many researchers lamented the fact that once their project was completed, there was no way to share the information they had collected. In fact, as an anecdote, two participants discussed their research on areas within the Chesapeake Bay. These researchers were literally working in areas within 100 miles of each other, but were working with different agencies and sources, and each spent much time tracking down, assembling and digitizing data. And each lamented that after doing all of their work, there was no easy place to deposit or share the new data resources they had created.
- Another challenge identified was the disparate repositories for the type of data needed to assemble a digital crust. The US Geological Survey has the most comprehensive data resources and has a goal of building a 3D geology for the nation. There are also state geological surveys with data (both digital and non-digitized) that could inform the digital crust. In fact, many participants at the workshop relayed developing relationships and working directly with state geological survey personnel, and employing graduate students to discover, digitize and format data from these offices. If we add in data being created, assembled and/or used by the geophysics and other geosciences communities, there is a clear need for governance and coordination.

TECHNICAL INFORMATION/ISSUES/CHALLENGES

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

- We envision a database composed of a collection of fundamental geologic units – including, but not limited to hydrostratigraphy and soil horizons. The system should accommodate these units as 3D GeoVolumes; this system should allow the size and shape of these geovolumes to evolve over time. We would envision a “reference” set of geovolumes, governed and maintained jointly by the academic community and the USGS that represent the consensus best available continental 3D geology.

- The system should be able to represent multiple interpretations or sets of geovolumes – a way to think about this might be the way geodatabases can have multiple layers that contain different representations or interpretations of surface properties. A user may come into the system through the continental 3D geology geovolumes but could then access other researchers’ interpretations of geovolumes over a specific area, find local or regional studies that have more detailed or high resolution information, etc.

- The system will need to contain and present substantial metadata in a way that allows both expert and non-expert uses to evaluate the interpretations and geovolumes for their quality, appropriateness, and fitness for use in different applications or models. *This was seen as a central, unresolved challenge by the workshop participants – communicating uncertainty, transferring the inherent knowledge, context and understanding of the scientist who makes the original interpretation, etc – are all key*

- The system must have an easy way for researcher to share and deposit their own data. The system must have ways for researchers to not only share their own data but to feedback to current data in the system – e.g. a researcher might contribute high resolution data set on a particular reason – this data should then be incorporated to our larger understanding of the system, and could/should result in a change in the "reference" set of geovolumes size and geometry over this area. This would require oversight/governance system to be set up.

- The system should have a way to represent and share proprietary or protected information (e.g. metadata only). Many researchers relayed experiences of working with data that is proprietary. Participants felt it was important that the digital crust convey the existence of this information as well as contact information for people to request access.

- Behind each geovolume requires a provenance, i.e., comprehensive archive of all supporting data and sources. Users could access this archive and work with the data directly to create their own geovolumes, extract data of interest, etc. The data system would have to accommodate variable resolution in x,y,z for the data underlying the geovolumes. The data system may have to accommodate gaps or "no-data" geovolumes.

- The domain of this data system would be from the land surface down to where data is available and material definable (a "goal" could be the brittle-ductile transition). The data system should easily integrate with other data systems as much as possible (e.g. surface data, DEMs, vegetation, etc. so that there are not mis-matches or discontinuities) so that researchers could easily assemble data needed to investigate critical zone or earth system processes.

- The data system should support a suite of data retrieval and analysis tools, allowing users to explore and access the data flexibly. Specific examples the workshop participants cited:
  - Flexible selection of spatial domain, grid resolution, generation of x-sections and geo-volumes
  - Enhanced visualization and ability to "video fly-thru" such as done by Google Earth; integration with other data sets. An example was given of viewing Google Earth or a DEM, and then having the ability to "peel back" the surface and see the subsurface underneath.
  - Algorithms to calculate grid cell properties (different means, std dev, functional forms, etc).
  - Ability to generate 3D grids of specific material properties (physical/mechanical, chemical, biological)
  - Ability to incorporate uncertainties or probabilities in 3D location. A specific example was researchers who wanted to create 3D GIS features of specific geologic features (e.g. sand bodies, areas of a specific threshold of an important property, e.g. high or low permeability) but wanted to be able to represent the uncertainty in the location of these features (since they are interpreted) – the system could create a 3D grid of probabilities of whether a feature was present, and 3D features that could represent specific probability thresholds as concentric shells.

- Although logical data models exist for representing 3-D geologic formations, the current tool set for working with 3-D geovolumes is inadequate to domain scientists. Standards for serving and exchanging such geovolumes are nascent at best.

COMMUNITY NEXT STEPS

1. **List of what your community needs to do next to move forward how it can use EarthCube to achieve those goals:**

    - Development of a generalized 3-D geologic map of North America is possible and would provide a useful starting point for developing the cyberinfrastructure necessary to maintain, evolve, and utilize the Digital Crust.
    -

    • Achieving that goal requires maintaining and growing the community of researchers working in this area and establishing some initial communications/community resources to share data and experiences. We can do this initially with the workshop participants and the Google Docs site to share research, resources, presentations, and thoughts. EarthCube could support these data communities through the Ning site, working groups, further workshops, to maintain momentum and coordination.

    • Expertise from current Earthcube groups (in particular, the groups looking at interoperability and semantics) could be utilized in working groups that could look more in depth at developing metadata models and standards to address the concerns of workshop participants given the uncertain, interpreted data products.

    • EarthCube will be needed to develop, implement and maintain the community governance needed to realize the digital crust as envisioned. Such a system would need to be jointly maintained by NSF (representing the academic community) and federal agencies (such as USGS).

EXECUTIVE SUMMARY: EARTHCUBE WORKSHOP RESULTS

(B. McElroy, Univ. of Wyoming: 12/11/12
L. Hsu, LDEO; W. Kim, Univ. of Texas; R. Martin, Univ. of Penn)

Earth Cube Workshop Title: Calling All Experimentalists- Experimental Stratigraphy

Introduction: Approximately 55-60 Earth-surface scientists gathered for a workshop held at the University of Texas (UT) at Austin over 12/11-12/12. The participants include groups of 18 from Japan, Korea and Europe and a group of 35 from the U.S., 20 of whom came from outside UT. Costs for the Japanese participants to travel to the workshop were leveraged from the JSPS funds they obtained separately. In addition the implementation of an online forum resulted in remote participation of another 7-10 scientists over the two days, including participants from Canada and Taiwan. Overall the size and diversity of the participating group played a substantial role in the success of this workshop.

Over the course of two days, participants carried out a set of community experiments, heard presentations from 7 keynote speakers, and held group discussions. The speakers specifically addressed the range of issues from their vision for the experimental Earth-surface science community, to recent scientific successes, to grand challenges that experimental science can address, and to community needs in pursuit of these goals.

The community experiments were conducted using input solicited from participants prior to the workshop. These experiments served as a focal point of discussion during the first day of the workshop and to facilitate openness during discussion through shared experience. In this effort partial experiments carried out by one subgroup were interpreted by another subgroup. This is all documented and publicly available through the workshop website (URL, https://sites.google.com/site/sedimentexperimentalists/workshop-experimental-stratigraphy). Beyond the results of the experiments, all other workshop materials, notes, and many presentations can also be accessed through the website.

Over the course of the workshop, three focused breakout discussions were led in addition to general, plenary discussion during experiments and keynote presentations. The break-out discussions were tied to each day's themes: day 1, current practices; day 2, current & future needs and best practices. The smaller discussion groups, 12-15 participants each, explored these themes for approximately one hour with notes taken. In addition, questionnaires were distributed to participants each day in both hardcopy and digital format. Along with the questionnaires, the presentations and discussions form the basis for the remainder of this document.

SCIENCE ISSUES AND CHALLENGES

Important challenges for advancing experimental stratigraphy: Participants identified several high-priority science issues that will be central to advances over the next 5-15 years

- How do we apply technical advances currently underway to experimental methods to create the next major advances in scientific knowledge? This will allow us to answer standing questions as well as ask completely new ones. These methods are likely to include:
  -Tomographic methods for the detailed *in-situ* investigation of strata as the evolve.
  -Long-range particle tracking methods for developing Lagrangian framework theories for sediment transport and deposition.

-Computational methods for measuring and modeling individual sediment grains in large, complex systems.

- What framework and model will allow us to gather and distribute large experimental data volumes for broad use beyond the original investigation? This is key to extracting greatest value from experimental data, increasing scientific efficiency at community level, and enhancing collaborations within and beyond the experimentalist community.

- How can directly coupling laboratory experiments to outcrop-based investigations accelerate advances in understanding? This approach is an excellent one for addressing major issues including:
  -Testing field-derived stratigraphic models (i.e. those directly tied to reservoir problems).
  -Addressing the grand challenge of integrating autostratigraphy and sequence stratigraphy.
  -Overcoming the community reluctance to incorporate experimentally-derived stratigraphic knowledge into stratigraphic models.

Current challenges to high-impact, Earth-surface science applications: Several themes emerged as consistent challenges faced within the discipline and its application to other disciplines

- How can we harness the collective abilities of the experimental community to conduct focused, timely experiments that would result in accelerated ability to answer large-scale questions. This model essentially divides up "parameter space" between various experimental facilities such that all conduct part of a single set of experiments together. The clear trade-off between number of participating facilities and time to complete experiments would allow for quick return on investment from many investigators.
- Will searching for general properties of sediment transport systems and stratigraphy move the community beyond applications of scientific knowledge by specific environment (e.g. delta, etc.? add a few examples? depositional versus erosive; subaerial versus subaqueous)?

- Can we specifically focus on remotely sensed data (e.g. image, topography, composition) for Earth landscapes and seascapes to provide tools for interpretation of these types of data on other planetary bodies. Remote sensing of environments and surfaces of extraterrestrial bodies is rapidly growing, and developing connections to exploration beyond Earth will be greatly aided by complete embrace of these data types for terrestrial systems.

TECHNICAL INFORMATION/ISSUES/CHALLENGES

Current and future needs for pursuing key science questions with brief elaboration:

- Many of the identified needs and challenges bridge the gap between scientifically technical issues and cultural issues of our scientific community. While many technical issues could be addressed at the individual investigator level, a community-scale effort would likely result in greater efficiency, and it is paramount for creating lasting cultural solutions. Most all needs identified below fit into this framework.

- Cultural: Difficulties with incentives for data sharing
  -Intensive in time and financial resources to produce an effective data sharing platform
  -No rewards for this kind of investment. Possible reward types could include

-Institutional support, i.e. recognition during tenure process
-Incentives from NSF for public data availability or reuse
-Recognition for data generation distinguished from interpretation in literature
-For long-term monitoring a funded investigator should still get to work exclusively with new data on interpretations before sharing
-Lack of long term solution disincentivizes investment in resource
-Scientists spend time on science and not on management

- Technical: Need for expertise in data issues within our community
  -No expert resources to call on for guidance and assistance in management
  -Need training for data management for students, etc. from the beginning of the project
  -Many institutions do not offer IT support to investigators

- Considerations for international cooperation
  -Language is a problem. For example, programming comments all in Japanese may be easier for the individual Japanese investigator but may create challenges in sharing code.
  -International agreement on sharing of data is an issue.
  -Coordination of physical and financial resources for hosting data is an issue.

- Opportunities to put our discussed ideas into action and test
  -Testbed data sharing site could be served distributively with single front-end combining data, metadata, models, etc.
  -Trial solution for linking documentation with data
  -Individual investigators can work independently to test differing solutions allowing faster discovery of models that do not work well. Compels move to more open source solutions.
  -Allows for growth of merit-based solutions in data/metadata structuring, i.e. not prescribed by committee but determined by acceptance and use
  -Opportunity to test if this expedites secondary use of data to answer broader scientific questions
  -More funded community discussions to further advance a plan for data storage and dissemination . ie: We have not gotten it all done in two days but current ideas can be verified or discarded.

- A funding model to make the dreams work - what would the funding model be? Institutional support / national agency / scientific organization

SCIENCE SCENARIO EXAMPLE

The following scenario is a typical example of a laboratory experiment in stratigraphy- one that attempts to understand process controls on coastal stratigraphic construction through manipulation of boundary conditions. This is a generalization of the group experiment that was run during the workshop. Similar experiments, specifically those incorporating sea level rise / fall, changes in sediment supply, and/or differential subsidence are common today because they relate to such diverse issues as coastal change, natural hazards, and natural resource exploration.

Experimental inputs and data types would include mass fluxes of fluids and sediments, normally these are variables for which a time series is recorded. The other boundary conditions would likely include a “sea level” curve and subsidence pattern. Again both being time series datasets and the subsidence would have a component of spatial variation and could also be represented as an evolving contour map through time. Other relevant data would include information about granular hydrodynamics, geometry of the experimental apparatus, and metadata regarding the details of the setup that would normally not be reported beyond a qualitative fashion. In order to quantitatively relate the boundary conditions to the resultant stratigraphic condition a range of observations might be made during and subsequent to the experiment. These might include repeat surveys of surface topography and bathymetry. Like subsidence, this is appropriately conceptualized as an evolving map. Time-lapse photography and derivative digital data such as shoreline position or distribution of channels or sediment type are often collected. Being directly derived from time-dependent spatial data, these types would have equivalent natures. Finally, the interests in stratigraphy itself compels a 3D spatial data that could be arrived at through physical dissection of deposits or through various types of remote sensing such as ultrasonic pseudo-seismic. These data could also be collected through time resulting in a 4D manifestation of a single stratigraphic property (e.g. grain size) (sliced deposit sections following completion of the experiment).

It is worth pointing out that methodologies for experiments of this type is diverse enough that it could have substantial impacts on the data recorded. Depending on technologies used, there is a direct trade-off between quantity of information and ability to record it. While those facilities that use computer-based control of boundary conditions can often record with greater temporal sampling rates, those that use manual controls are often much more limited in the data volumes associated with boundary condition changes. Similarly, in the latter, all notes (both inputs and observations) would likely be recorded by hand in participant notebook while the former is recorded digitally during the control process. Example data from our group experiment can be found through the workshop website (URL, https://sites.google.com/site/sedimentexperimentalists/workshop-experimental-stratigraphy).

EXECUTIVE SUMMARY: EARTHCUBE REAL-TIME WORKSHOP RESULTS

June 17 and 18th, 2013
Boulder Colorado

Organizing Committee:
Mike Daniels, NCAR (Chair)
V. Chandrasekar, Colorado State University
Sara Graves and Sandra Harper, University of Alabama - Huntsville
Branko Kerkez, University of Michigan
Frank Vernon, Scripps Institution of Oceanography/University of California - San Diego

Workshop Breakout Teams:
Tim Ahern, Incorporated Research Institutions for Seismology
Jennifer Arrigo, CUAHSI
Janet Fredericks, Martha's Vineyard Coastal Observatory/Woods Hole Oceanographic Institution
Alexandria Johnson, Purdue University
Kate Keahey, Argonne National Laboratory / University of Chicago
Charlie Martin, NCAR
Jim Moore, NCAR
Mohan Ramamurthy, UCAR Office of Programs
Siri Jodha Singh Khalsa, NSIDC
Greg Stossmeister, NCAR

Earth Cube Workshop Title: Integrating Real-time Data into the EarthCube Framework

Introduction: The primary findings of this workshop contend that real-time data have the potential to transform the geosciences by enabling adaptive, feedback-driven experimentation and improved societal decision making. Once the cyberinfrastructure is in place, the ability to act on and analyze data as it is collected will enable the discovery of scientific phenomena that may otherwise go unobserved and unexplained. Specifically, instantaneous feedback from sensing devices will enable real-time hypothesis testing, improving the quality of ongoing interdisciplinary experiments through adaptive reaction, while facilitating unaliased observations of space/time scales. Continuous, real-time data will lead to new, potentially unanticipated data discoveries as scientists respond to emerging, in-situ phenomena. Furthermore, real-time data is relevant and engaging to both scientists and the public. The utility of real-time data to societal decision makers cannot be understated, as it will enable new suites of operational support and disaster mitigation systems.

There exists a significant opportunity to realize this potential through improved sensors, better networks, advanced algorithmic techniques, and on-demand availability of computer resources. A concerted effort is required across the geoscience and cyber infrastructure communities to define the unique nature of real-time data streams and their role in the future of *EarthCube* and broader NSF initiatives. In summary, the integration and development of real-time capabilities will have significant transformative effects on Earth science in the next five to fifteen years.

The EarthCube real-time data workshop took place June 17th and 18th in Boulder, Colorado, and involved 76 participants from a variety of state and federal agencies, academic institutions and industry. A broad spectrum of geosciences was represented, including but not limited to Hydrology, Oceanography, as well as the Earth, Atmospheric, Space, Polar and Cyberinfrastructure sciences. To motivate breakout sessions and to provide use-case ideas, prominent experts from these domains presented real-world examples of the need for real-time data in their experimental campaigns. The major scientific and technical challenges behind the integration of real-time data into the EarthCube framework, as well as motivating use case scenarios were outlined by the workshop participants and are summarized below.

SCIENCE ISSUES AND CHALLENGES

1. **Important science drivers and challenges:** Participants identified several high-priority science questions that will be the focus of interdisciplinary efforts during the next 5-15 years (list 3 to 6).

    - How can we better use real-time data to understand the processes of high impact events or phenomenon and translate that knowledge to better response procedures? Examples of critical cases include, but are not limited to:
        - Improved hurricane track and intensity forecasting; prediction and response to coastal inundation and shoreline breaches
        - Better understanding of tornado and severe convective storm genesis and warning
        - Earthquake and tsunami prediction
        - Better understanding, predicting, and managing of Hydrologic Extremes, e.g. flash floods
        - Early detection of harmful algae blooms
        - Prediction of large solar flare events for assessment of damage satellites.

    - How can we better understand scientifically compelling phenomenon with adaptive real-time, feedback-driven science? The following strategies optimize the scientific values of our measurements and enable new discoveries:
        - Dynamic sampling strategy to collect, analyze, and respond to real-time data
        - Response examples: Moving platforms, changing scan strategies, adjusting flight patterns, automatic adjusted of instrument signal processes, deployment of additional instruments, etc.
        - Using models in conjunction with adaptive strategies to improve sampling
        - Real-time awareness of instrument status to support rapid response to issues and improve data quality
        - Instrument validation (is it responding to its environment and can we adjust the instruments to improve the response?)
        - Tools that enable broad communication and collaboration during real-time mission oriented research
        - Detection and discovery of new, unexpected phenomena that need to be explored further
        - Tracking and sampling of transient phenomena

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).

- Interoperable streaming protocols and metadata (including consistent and accurate time stamping and spatial coverage) for real-time data streams across the geosciences domain do not exist.

- There are few, if any, mechanisms and processes in place to assess the quality of real-time geosciences data.

- Visualization tools for interdisciplinary real-time data of varying spatial and temporal coverage need to be developed.

- Valuable real-time data streams are often not integrated with downstream decision support systems used by emergency managers, etc.

- Real-time data streams need better connections to prediction models and/or systems that produced derived products.

- The scientific community generally does not properly address real-time data at the same level of archival data or Big Data in terms of data management plans and other data-focused initiatives.

TECHNICAL INFORMATION/ISSUES/CHALLENGES

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

- Improved community infrastructure: access to improved communication infrastructure, on-demand computing and protocols for data exchange

- Metadata generation for real-time data streams and tracking of provenance

- Real-time signal processing, calibration, and quality control: existence of standardized software libraries

- Real-time computing: software that provides the ability to process, produce, and transmit derived products in real time.

- Tools for integrating and assimilating real-time observations: from differing geospatial and temporal resolutions

- Playback tools for re-creation and analysis of phenomena and the observed environment of past experiments

- Frameworks and secure mechanisms for remote operation of instruments

- Real-time visualization of observations made at different temporal and spatial scales

- Data discovery and access including data subsetting of large bandwidth streams
- Rendering of observations with widely different time scales for real-time displays
- Decision support tools and integration with tools for emergency management
- Engine/middleware/platform that will combine all these capabilities for the community
- Developing networks for dissemination - including social media, apps, and user driven interfaces/portals including citizen science, crowd sourcing and open data access
- Mechanisms to discover software and hardware for real-time acquisition and processing and to provide guidelines in implementing real-time capabilities (eg, SUB/PUB real-time streams, buffering data for remote access, etc), education

COMMUNITY NEXT STEPS

1. **List of what your community needs to do next to move forward how it can use EarthCube to achieve those goals:**

- Community Development:
  - Assess system interoperability between geosciences real-time data providers and users
  - Establish best practices with scientists and CI experts across domains engaged in real-time data systems
  - Share knowledge, tools and approaches to real-time data experts across the geosciences
  - Refine requirements needed for real-time data streams to connect to downstream decision-making tools and processes
  - Increase awareness of the real-time data streams that are in existence among the geosciences community to facilitate new uses of these data

- Prototyping:
  - Pilot projects, demonstration testbeds, identification and development of real-time data capabilities
  - Build real-time stream translation tools across geosciences disciplines
  - Develop of a prototype framework for real-time control of instruments that can be more generally applied to the geosciences
  - Respond to missing capabilities in Section 2 above such as real-time quality control mechanisms, real-time metadata standards or real-time visualization tools that span geosciences data of varying spatial and temporal domains
  - Begin work to develop a “universal real-time infrastructure” where data streams are captured, organized and made quickly available, making it much more likely that they will be adopted by stakeholders who have noticed a new phenomena and want to examine it in the context of current events

- Capacity Building:

- Work with undergraduate and graduate students to engage them and popularize real-time science and data
- Develop the next generation workforce by exposing students to real-time data and its importance to science
- Explore and engage the private sector to meet the collective needs of the real-time geosciences data community

DRAFT
EXECUTIVE SUMMARY: EARTHCUBE WORKSHOP RESULTS

April 24-26, 2013. Millennium Harvest House Hotel, Boulder, CO

Conveners: Albert Kettner (Univ. Colorado) and Emilio Mayorga (Univ. Washington)
Organizing Committee: Anthony Aufdenkampe (Stroud Water Res. Center), Anne Carey (Ohio State Univ.), Basil Gomez (Univ. of Hawaii at Manoa / State of Hawaii Commission on Water Resource Manag.), Laodong Guo (Univ. Wisconsin Milwaukee), Sherri Johnson (Oregon State Univ. / US Forest Service), Bernhard Peucker-Ehrenbrink (WHOI), Peter Raymond (Yale)

Earth Cube Workshop Title: *An EarthCube Domain Workshop integrating the inland-waters geochemistry, biogeochemistry and fluvial sedimentology communities*

Introduction: This 2.5-day workshop brought together 55 diverse participants (with half being early-career: about 9 graduate students, 9 postdocs and 10 assistant-level faculty or research staff) and an additional 10-15 virtual participants to discuss the status, needs and opportunities regarding cyberinfrastructure impacts on the advancement of overlapping communities of scientists that address sources, composition, processes, fluxes and fates of constituents in terrestrial surface waters. 82 individuals registered for the workshop (including virtual), but several were unable to attend due to last-minute conflicts or health issues. Participant recruitment placed a special emphasis on inviting and attracting domain scientists, including ones with strong interests and activities in data integration and synthesis and cyberinfrastructure-enabled science. The workshop was structured to enable substantial exchange on scientific results and interests in order to both stimulate breakout EarthCube discussions and maintain strong participant engagement; these exchanges were facilitated through keynote presentations, posters, and many brief (2-5 min) "pop-up" presentations on topics spanning the range of disciplines, including existing cyberinfrastructure serving the represented disciplines. Participants were divided into breakout groups according to research approaches (scale, environments, disciplines), and re-mixed across breakouts to maximize the generation of new perspectives.

A goal of the workshop was to help define and form a community of aquatic scientists who have common ground and overlapping interests in their study of inland waters geochemistry, biogeochemistry and sedimentology, but may communicate rarely due to disciplinary fragmentation or regionally specific interests. The focus on a unifying set of environments or "hydroscapes" (dynamic water systems) and data needs and conceptual frameworks involved when addressing terrestrial, aquatic and atmospheric and anthropogenic influences on surface waters indeed proved to be an effective mechanism for identifying common data and cyberinfrastructure needs across disciplines.

Workshop participants identified important scientific drivers or grand challenges for advancing cyberinfrastructure capabilities benefiting their disciplines, as well as technical and other impediments to addressing these scientific drivers. They compiled an extensive list of data, software, tools and modeling resources used by the community, and created a list of recommendations and consensus on three unifying, grand-challenge use cases for advancing and leveraging cyberinfrastructure capabilities.

SCIENCE ISSUES AND CHALLENGES

1. Key science drivers and challenges: The study of dissolved and particulate matter is of relevance to geoscientists and ecologists and encompasses diverse landscape scales and types, element and material cycles, approaches, and data collection contexts. This broad community is highly interdisciplinary; two different breakout groups came to the similar conclusion that this interdisciplinary nature makes it very difficult to label data sets as being "within" vs "outside" the discipline. The sub-disciplines are complementary and interact with one another. Several unifying themes emerged, containing more detailed questions and challenges:

- We are in the era of anthropogenic changes.
  - Need for understanding current states in relation to historic mechanisms driving the systems. The role of *legacies:* Climate history, soil structure, past disturbance, past land use.
  - What is the magnitude of climate change impacts vs. direct human perturbations such as land use change, aquatic environment modifications, and hydraulic engineering?
  - What are the global trends in carbon export, concentrations, gas evasion fluxes, and burial?
  - How will climate change affect higher latitude changes/creation of wetlands?
  - Trajectories and impacts of wetland degradation and restoration.
  - Advancing understanding of water ecosystem services to address landscape management.
- Connectivity: Lateral linkages via water transport.
  - When and where do hillslope flowpaths connect and disconnect?
  - What is the impact of groundwater connectivity on stream processes?
  - How do we connect flowpaths and systems across scales?
- Temporal perspectives: Predicting time of response to climate change and disturbance across biogeochemical response variables and water body types.
  - Pulsed events, extreme events (hurricanes, land slides, …).
  - How do seasonality, magnitude and duration of events influence biotic responses?
  - How does temporal variability impact societal needs or benefits?
- Spatial perspectives: Predicting zones of conservation, transformation, propagation within inland water networks across range of response variables.
  - Defining and mapping the time-varying *hydroscape*, including small streams.
  - Upscaling different systems and fluxes to regional, continental, and global scales.
- Grand goal: Integrating and translating across spatial scales and forecasting in time. Needs:
  - Improving the mechanistic understanding of processes.
  - Increasing spatial and temporal extent and resolution of observations.
  - Dynamics of fluxes, process rates, and system scales.
  - Linking across different types of processes and forcings: physical, biological, chemical, geomorphic, and anthropogenic.
  - Linking understanding of quantity and composition of complex constituent mixtures.
  - Using fine-scale data and understanding to inform global scale understanding.
  - Determine how water, sediment and biogeochemical fluxes *throughout* a river basin are connected and affected by event magnitude, duration, sequencing and spatial extent.
- Other challenges:
  - How do relationships between discharge and concentration impact downstream ecosystem function, and how do food webs impact biogeochemical and even hydrologic responses?
  - Estimate global time series of monthly carbon burial fluxes in all aquatic depocenters, and carbon gas exchange fluxes across all water surfaces.
  - How do floodplains function geochemically and geomorphologically?
  - Understanding delta subsidence and retreat due to decreased sediment supplies.

TECHNICAL INFORMATION/ISSUES/CHALLENGES

Current challenges to high-impact, interdisciplinary science:

- Addressing the various types of data: At a point (across space, these will become); Spatial (local regional global); Temporal (minutes, days, weeks/months, annual, decadal, geologic scale); Field Samples; Modeled Samples
- Datasets that bridge measures of quantity and composition of complex constituent mixtures
- Disconnect between continuous measurements and concentration data.

- Challenge of observing and characterizing hot spots and hot moments (fluxes and processes that are highly concentrated spatially and temporally)
- Using, sharing, and coupling broader models (geochemistry, hydrology, etc)
- Challenges finding data.
  - Lots of free data, very different formatting, provenance, other data characteristics.
  - Zero order challenge is knowing what data is out there, knowing what resources provide access, etc. Data discovery is lacking.
  - Challenge in finding linked data - spatially and temporally
  - Downscaling and upscaling data
  - What are the core capabilities that end-user domain scientists need in terms of data management/cyber-infrastructure?
  - Can we filter content based on assumed associated data?
  - Community standards for data management would make our science practices more efficient. Low transaction cost is necessary for adoption.
  - Need benchmark data (such as for training/validating models). Model intercomparison portals, testing our model quality.
- Challenges using data.
  - Data quality varies, across soil types, DEM, sensors precision, accuracy
  - Enormous datasets that are difficult to use
  - Units are different, and metadata doesn't always provide clarity
  - Curation. Heterogeneous data quality. Lack of information about quality. Reviews of data sets? Summary of data quality and characteristics?
  - Clearinghouse function of Earthcube? Meta DataBase
  - Vocabulary variation. Semantic search.
- What are the pressure points for the community? AGU and Nature Geo saying "we won't accept this unless you link your publication to data". But...where does it go? Let's go to a couple of the heavy hitters and ask them to be the bad guys. Others will follow.
- Leverage other, related disciplines for their solutions to similar problems. Which atmospheric/ oceanic lessons would be translated to our domain?
- Need to change culture about code sharing. What are the incentives? Why do we have different rules for code vs data sharing?
- How do we decide the scale for making decisions? E.g., OGC deciding on standards vs grassroots community?

1. Desired tools, databases, etc. needed for pursuing key science questions.

- National data set from water treatment plants and sewage treatment plants (they generally are apprehensive and don't share but have great data)
- International data sets (some countries do not share)
- Smoothed county level data
- Improved and standardized statistical approaches for small systems
- High resolution data throughout the hydrological cycle, not just field season campaigns
- communication about current projects, research activities
- Tools for coupling models are important and useful.
- Rating datasets and models. Need to understand relative value of data and models faster.
- Need standards for data exchange and formats.
- Understanding and curating what has been done and what could be available.
- Digitizing the wealth of information that exists behind us (historical data).
- A large, comprehensive catalog? Consistent formatting. Need crosswalk for vocab.
- Central retrieval system; centralized searching, not necessarily hosting data physically

- We need mobile science apps to make field work more efficient. Improved models of concentration discharge relationships (*how do we share models?)*
- Maps of built infrastructure information and data (tile drainage, pipeline, sewage treatment outlets, past land use)
- Ground water chemistry database
- Continuous categories of soil maps and soil chemistry
- Fertilizer use data
- Watershed activity for research
- Historical maps of land use, lead deposition etc
- Species distributions of fish, invertebrates, amphibians, native, invasive species
- Hyporheic flow paths
- Soil moisture maps

COMMUNITY NEXT STEPS
1. List of what your community needs to do next to move forward, and how it can use EarthCube to achieve those goals.

EarthCube activities.

- Sign up and participate in EarthCube Ning site group. Send email to all participants, all who registered for workshop, and otheres who expressed interest, to join this EarthCube Ning group.
- Make the Workshop Google Drive accessible from web site via link that gives read-only access; and send invitation to participants to give individuals edit access.
- Participate in second RCN call, if there is one.

Community building. Maintain discussions and momentum

- Continued at least informal gatherings at conferences, including AGU; and later sponsor a special session at the Joint Aquatic Sciences Meeting in May 2014
- Identify potential sources of support for subsequent meetings, and general alignment
  - US: SESYNC / Water Science Software Institute; CUAHSI, CZO, LTER StreamChemDB, USGS CIDA, Global Rivers Observatory
  - International: GEOSS/GEO Water Quality Community of Practice, IAEA, GLORICH, GEMS/Water, RECCAP
  - Relevant observatories: GLEON, NEON, CZO, LTER/USFS
- Prioritize focal areas - ask specific questions (eg, Eutrophication; Human influence on reservoirs, lakes and wetlands; Mechanisms that control gas exchange). Success of past efforts has been in creating a database that handled narrow sets of data.
- Identify low-hanging-fruit cyberinfrastructure steps and products, with current funding:
  - Increased integration with terrestrial work
  - Model sharing
  - Explain relevance to societal problems.
  - Get universities involved. Graduate training - consider a piece of the planet; do a synthesis of the available data; visit, tour, and meet with researchers. Model this after the field-course program of the Duke-university based Organization for Tropical Studies.
  - Database building
    - Divide "wish list" into efforts that we need other communities to tackle vs. efforts that we need to take on
    - Create an interactive wiki with data sources, searchable per region, use/parameter group
    - Online training webinars
      - Loading/using data in CUAHSI HIS and Water Data Center
      - IEDA System for Earth Sample Registration (SESAR), http://www.geosamples.org/
      - IEDA EarthChem Data Library, http://www.earthchem.org/library

    - Begin loading our dark data in these systems, even if not perfect for our community.
    - Assemble a priority list of dark data (e.g., Hans Eugster and Peter Kilham saline-lakes datasets).
    - Develop/distribute checklist for desirable metadata, templates for data
- Organize and sponsor Software Carpentry “boot camps” for this community
- Write white Paper, for Eos or peer-reviewed journal

Advance the three Consensus Use Cases, both with EarthCube and more broadly.

1. **Title:** *GoogleEarth-like "H2O" (Headwater to Ocean) Data/Model Access and Visualization.* **Goal:** Create a portal that provides access to constituents transported by rivers (from headwater to the coastal ocean) to better constrain fluxes and the understanding of processes that determine fluxes. This includes sensor data (in-stream, remote sensing), legacy data (links to existing data repositories, but also mining/rescue of dark data), model output, integrated spatial data characterizing watersheds, and careful quantification and propagation of uncertainties.
2. **Title:** *The role of "events" on water, temperature, sediments, solutes and ecology: comparing case studies of ENSO impacts ("IMENSO", IMpact of ENSO) on the flux of sediments and aquatic biogeochemistry.* **Goal:** Examine the role of ENSO climate variability on sediment/ carbon/ nutrient fluxes through case studies of data rich and data poor watersheds around the world (e.g. California, New Zealand, Amazonia, Ethiopia, Australia).
3. **Title:** *Role of inland waters in historic and contemporary global biogeochemical cycles (GBC).* **Goal:** Develop an understanding of inland waters role in GBC, based on an understanding of water quantity (fluxes, storage, residence times) as a foundation to understanding carbon and nutrient fluxes, with an emphasis on greenhouse gases. Begin with a contemporary, process-based global baseline understanding, then predict past and future.

EXECUTIVE SUMMARY: WORKSHOP RESULTS

(E. Hajeck, Penn State: 8/8/12)

Earth Cube Workshop Title: MYRES V: The Sedimentary Record of Landscape Dynamics

Introduction: Meetings of Young Researchers in Earth Science (MYRES) is a community-driven initiative aimed at promoting interdisciplinary research efforts among early-career scientists from across the world. MYRES V: The Sedimentary Record of Landscape Dynamics brought together a wide range of early-career geomorphologists, sedimentologists, stratigraphers, and geodynamicists interested in bridging Earth-surface and solid-Earth research in order to better understand the evolution of Earth's environments over a range of temporal and spatial scales, and in response to a variety of tectonic and climatic forcing events.

Workshop participants (54 in total) were dominantly from the US (80%) represented a variety of disciplines including geomorphology (32%), sedimentology (32%), stratigraphy (27%), and geodynamics (9%). Main outcomes of the EarthCube workshop discussions are summarized below.

SCIENCE ISSUES AND CHALLENGES

1. **Important science drivers and challenges:** Participants identified several high-priority science questions that will be the focus of interdisciplinary efforts during the next 5-15 years.
   - What processes are relevant to understanding landscapes and mass flux (i.e. sediment budgets) in the past, present, and future across different temporal and spatial scales?
   - How is sediment generated and changed as it moves through the landscape?
   - How does downstream transmission of Earth-surface materials filter and record the frequency and magnitude of Earth's environmental changes?
   - How does life influence surface processes and transform environmental signals preserved in the sedimentary archive?
   - To what extent do extreme events control landscape evolution and stratigraphy?
   - How do the effects of tectonic and climate conditions propagate through the landscape and depositional system? At what scales?

2. **Current challenges to high-impact, interdisciplinary science:** Several themes emerged as consistent challenges faced across disciplines.
   - Many researchers approach data collection and data comparison individually, compiling datasets for use in their own research group. Often this includes tracking down other researchers willing to share data, digitizing legacy data from older publications, and spending tremendous amounts of time looking for proper metadata and checking data quality. A community effort to make data available for download, along with thorough metadata would save researchers tremendous amounts of time.
   - It is currently difficult to integrate across disparate datasets (e.g., field, experimental, and modeling data) and across disciplines (e.g., geophysics, atmospheric science, oceanography, etc.). Existing databases and online modeling resources often require a high level of insider

knowledge for the resources to be fully utilized; this is a barrier to entry for researchers trying to collaborate from other disciplines.

- Inconsistent formatting, file types and metadata make compiling interdisciplinary datasets difficult.
- The culture of collaborative science in which data are openly and easily shared is only just being established. Many researchers were not trained to collect and report data in a format that would be usable for other researchers (particularly those they do not know), and currently it can be very difficult to access others' data (this is, perhaps, particularly true of field observations). NSF's data-gathering requirements are an opportunity to establish a new framework and new vehicles for data sharing among researchers.

TECHNICAL INFORMATION/ISSUES/CHALLENGES

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

- There was a quickly realized consensus that a Google-Earth-like data clearinghouse would be tremendously helpful. This would be a place where existing community datasets could be searched both by topic/index and geographically (as well as temporally, for historical and stratigraphic data). (For example a search for "suspended sediment" and "discharge" might return datasets from the USGS, the Army Corps of Engineers, individual PIs, local/state and international agencies.) Ideally, once desirable datasets are identified, a researcher could then download them in a similar file format/structure. Physical and numerical modeling results could also be included and geographically cross-referenced by the lab of origin and a specific location (if a model were related to a field case, for example), and model codes could be shared (in a similar manner to what is currently done through CSDMS).
- Participants agreed that Google Earth (or similar intuitive geospatial interface) itself would be a desirable backdrop for this type of community resource and there was no need to reinvent the wheel in terms of user interface, for example. Access to linked datasets via a large, searchable data clearinghouse (where file downloading and metadata storage were reasonably uniform) would also help improve access and usability of disparate datasets. This might mean, for example, that a researcher would only need to learn one data upload/download system, which would empower users to access geological, geophysical, biological, and climatology data, for example, via the same interface, rather than having to learn a new protocol to access data from each discipline.
- A centralized data clearinghouse would also provide a place for PIs collecting new data to upload their results, thereby blending both existing databases and accommodating the needs of researchers who currently rely on ad-hoc arrangements to store and share their data. Although NSF's new data-sharing requirements are separate from EarthCube, participants expressed concern that if new data acquisition/sharing isn't incorporated into the EarthCube model, some of the problems and challenges listed in section 2 will persist.
- There was also strong interest in the suggestion that resources be allocated to digitizing and updating legacy data that is not currently available in digital form. Participants were very enthusiastic about this idea and felt that it would be a high-yielding investment.
- The "universal Earth-science database" concept generated the most excitement among participants. Participants were less concerned with visualization and modeling resources, in part because it seems that existing software and collaborative websites (e.g., CSDMS) are

suitable for accomplishing important research goals (or at least are not viewed as significant barriers to progress), although improvements in visualization software and access to expensive software licenses (particularly for evaluating LIDAR and seismic data) would be helpful. Ultimately, challenges locating, accessing, formatting, and compiling data are currently frustrating and stifling to participants in this workshop.

EXECUTIVE SUMMARY: EARTHCUBE WORKSHOP RESULTS
Catalina Island, USC Marine Station August 20-23, 2013
Final Draft: Ed DeLong and Workshop Participants, August 29, 2014

Earth Cube Workshop Title: Ocean 'omics science and technology cyberinfrastructure : current challenges and future requirements.

Introduction: The overall goal of this EarthCube workshop was to bring together a group of leaders in ocean 'omic science and computer science, to help identify and prioritize a set of unifying scientific drivers and cyberinfrastructure requirements necessary to enable the storage, curation, federation, and comparative analyses of large and small scale ocean 'omic datasets, that are emerging from many recent scientific efforts. Applications of these data have great potential for improving our understanding ecosystem processes and predicting their future trajectories, but the necessary computational tools for doing so are still lacking.

A large group of ocean scientists and oceanographers are now employing 'omics approaches to characterize and quantify the nature, distribution and function of organisms in ocean ecosystems. "Omics" is defined here as the collective molecular or biochemical characterization of pools of biological molecules, such as genes and genomes, transcripts and transcriptomes, proteins and proteomes, and small molecules, metabolites and metabolomes, that together encode the structure, function, dynamics and activities of an organism or organisms. The tools and datasets that encompass 'omics science are diverse, complex, and rapidly expanding, and require the construction, curation, and query of diverse federated databases, as well as development of shared interoperable, "big-data capable" analytical tools.

To achieve the workshop goals, participants (46 in total, predominantly U.S. citizens) represented a variety of relevant disciplines including microbial oceanography and genomics (35%), phytoplankton genomics and ecology (15%), deep-sea microbiology (24 %), cyberinfrastructure and genomic scientists (15%) and Foundation representatives (11%). The condensed outcomes of Ocean Omics EarthCube workshop discussions are summarized below. A more detailed report will be provided after the vetting of comments and breakout group summaries with the workshop attendees.

SCIENCE ISSUES AND CHALLENGES

1. **Important science drivers and challenges:** Participants identified several high-priority science questions that will be the focus of interdisciplinary efforts during the next 5-15 years.

   - How do physical and chemical oceanographic parameters and biological population structure and function co-vary within and between different oceanographic provinces? Do steep physical and chemical gradients result in steep microbial functional gradients and drive changes in microbial biodiversity? Do feedbacks exist in both directions?

- How does 'omic and population plasticity in microbes bolster ecosystem resilience to disturbances? How does global change and environmental disturbance impact genomic repertoires, transcriptional organization, protein and metabolome content, and biogeochemical activity?
- What are the underlying molecular and biochemical mechanisms that regulate the physiological responses of microbes to environmental change, and their downstream biogeochemical consequences and feedbacks?
- How do microbial communities in the ocean fluctuate as a function of distance from land, seafloor spreading centers, gyres, and upwelling zones? How do they change as a function of geochemistry, currents, and crustal age? How does this affect the flux of matter and energy in the surface and deep sea?
- By what microbially-mediated mechanisms does rapid polar climate change affect the budget of greenhouse gases in the context of permafrost thawing and dissolved organic carbon release and transport, in time and space?
- How can 'omics data be more effectively leveraged into predictive frameworks for understanding ecosystem processes and their future trajectories ? How can 'omics data be distilled into tools useful to managers and stakeholders for efficiently monitoring ecosystem change and detecting ecosystem impairment?

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

- It is still a challenge for the community to develop, validate and implement standardized and federated procedures for sample collection schemes, sample QC/QA, data formats, annotation workflows, and data analyses, and to integrate those with geochemical, biological, and physical oceanographic data over multiple nested spatiotemporal scales.
- The community currently has limited access, storage space, and transfer mechanisms for sharing and archiving of raw data, processed data, data products from workflows, and records of the provenance of data analyses.
- The community generally has limited access to large scale, high performance compute capabilities necessary for the annotation, comparison, statistical analyses and other workflows required for analyses of large scale ocean 'omic datasets.
- There are new non-sequence-based datatypes (e.g. mass spectrometry used in metabolomics) emerging that will need to be stored, accessed and analyzed and federated with other environmental and 'omic datastreams.

- The community lacks sufficient tools for simultaneous visualization and intercomparison of heterogeneous datatypes (e.g., environmental, 'omic and oceanographic datasets).
- It is currently difficult to integrate emerging 'omics datasets and analyses with existing and developing physical and biogeochemical models. This is partly an analytical problem (e.g., the mapping of genes and pathways onto their respective biogeochemical activities), and partly an integration problem, requiring the combination of quantitative 'omics-derived biogeochemical information, with quantitative geophysical and geochemical models.

TECHNICAL INFORMATION/ISSUES/CHALLENGES

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

- Omic database development is required for curation, maintenance and data standardization that will allow for easy data submission, extraction and query. As well, tools for rapid and simple data query and metadata association are necessary. This includes federation with non-sequence-based datasets (e.g. metabolomics and lipidomics) into existing/emerging oceanographic 'omics database/analysis/visualization platforms. Environmental 'omic databases need to be: (1) federated (i.e., all datasets are interoperably queriable and transparently accessible), (2) curated (validated and updated, as for example NCBI nr datasets), (3) sustained (i.e. a five-year commitment of support is not sufficient), and importantly, (4) intuitively accessible to a broad range of scientists, and the public.

- The ocean 'omics community would benefit from "Google-like" or "Kayak-like" search and suggestion functions/engines, that could query across complex and heterogeneous, federated environmental, oceanographic and 'omic databases.

- Tools and mechanisms are required for access to high performance computing and statistical analyses of large scale 'omic datasets, that could accommodate both naïve users as well as experienced "power users". One possibility is a user facility that functions similarly to UNOLS oceanographic facilities, that would provide access to software developers, bioinformaticians, and analytical tools, as well as the hardware required (storage facilities, servers, clouds, etc) required for 'omic analyses. Researchers could request access to this facility in association with successful grant applications, as with UNOLS. Extending the capabilities of BCO-DMO or similar services also seems another tractable model.
- Tools are required for more intuitive, accessible and integrated visualization of linked environmental, 'omic and oceanographic (and other interdisciplinary) data

sets. Statistical tools and techniques for dataset inter-comparison and spatiotemporal modeling also are critical and need further development.

- The community would benefit from access to a web clearing house/portal with links to standard “ocean 'omics” best practices, algorithms, software and workflows, as well as analytical and statistical methods under development, with entry points for both naïve and power users, would be a useful resource for the community.

COMMUNITY NEXT STEPS

1. **List of what your community needs to do next to move forward and how it can use EarthCube to achieve those goals:**

- Cross train and educate computer scientists and engineers, and ocean and earth scientists to improve communication and collaboration among disciplines. This includes training and education to develop cross-disciplinary expertise within and between bioinformatics, the Earth sciences, and the Ocean sciences.
- Facilitate access, availability and utilization of NSF supercomputers for the Earth and Ocean sciences communities. (Using government supercomputers should be as technically easy, and as feasible as accessing the Amazon EC2 grid).
- Plan and initiate a community Research Coordination Network to support cyberinfrastructure technology and infrastructure development and education in ocean 'omics.
- Promote the development of an EarthCube system that would combine the facilitative role of the BCO-DMO database (or similar), with novel and flexible analyses and visualization services for analyzing and exploring ocean omics oceanographic data (e.g., Ocean Data View-like software and tools, for ocean ‘omics data).
- Further identify ocean 'omics cyberinfrastructure “parts” (e.g. dataset curators, search engines, high performance compute facilities, workflows, user analytical facilities, developers, etc.) that are operational and in use now, and determine which ones might be further improved, developed, federated, and networked into a functional EarthCube community ocean 'omics cyberinfrastructure solution.

EXECUTIVE SUMMARY: DEFORM & COMPRES EARTHCUBE WORKSHOP RESULTS

(Chris Marone, Jay Bass, Przemyslaw Dera, Heather Savage,
Tom Duffy, Terry Tullis and workshop participants)

Earth Cube Workshop Title:

EarthCube End-User Domain Workshop for Rock Deformation and Mineral Physics Research, Nov 12-14, 2013

Introduction:

Workshop participants addressed the current state, future challenges, needs, opportunities, and directions of cyberinfrastructure as related to research in rock deformation (DEFORM) and mineral physics (COMPRES). The workshop leveraged the high degree of compatibility that exists between the DEFORM and COMPRES communities. A key goal of the workshop was to identify scientifically transformative activities that could be facilitated by EarthCube.

A total of 76 participants gathered for 2.5 days, including 18 pre-tenure faculty, postdocs and other early career scientists. Workshop participants represented a variety of disciplines including mineral physics (36%), rock mechanics (34%), program managers from NSF/DOE/USGS (11%), cyber-science and engineering (9%), structural geology (5%), and geodynamics (5%). The agenda included 15 keynote talks, 12 lightning talks, and 24 posters presentations during an evening session. The workshop featured vigorous discussion from every participant during three plenary sessions, three breakout sessions, 8.5 hours of scheduled, free form discussion, and 9 hours of informal discussion. Outcomes of the workshop are summarized below.

SCIENCE ISSUES AND CHALLENGES

1. **Important science drivers and challenges:** Participants identified several high-priority science questions that will be the focus of interdisciplinary efforts during the next 5-15 years.

- As planets age and cool, how do their physical & chemical properties and internal structures evolve under the extreme conditions of pressure and temperature? What material transformations occur in complex, multiphase systems within planetary interiors and how do these impact key compositional and rheological boundaries such as the lithosphere asthenosphere boundary and the D'' region at the base of Earth's mantle?
- What processes determine where earthquakes occur to define the seismogenic zone, and how do they influence the tsunami-generating potential of seismic rupture at subduction megathrusts? What are the factors that dictate the spectrum of fault slip behaviors and the physics of slow earthquakes where self-sustained, quasi-dynamic ruptures propagate at velocities dictated by unknown processes.
- How do the physical and chemical properties of planetary materials control the dynamics and magnetic behavior of Earth and other planets?
- How can we best utilize seismological data and models from EarthScope and other sources to determine the composition, temperature, and flow fields that produce tectonic processes on Earth's surface.

- What are the factors that determine the brittle ductile transition within Earth's lithosphere and how does the transition from seismic to aseismic slip vary with strain rate? How can we image the slip distribution of large crustal earthquakes, for example on the San Andreas fault, to illuminate the properties of the deep crust, the brittle ductile transition, and the rheology of the lithosphere?
- How do microstructures evolve at high strain and what feedbacks connect this evolution to deformation, seismic properties, and fluid transport processes in Earth's lithosphere? How do experimental results inform interpretation of field data/observations and vice versa? We need to develop robust flow laws for multi-phase materials and to advance our understanding of anisotropic viscosity for Earth materials.
- Many socioeconomic, environmental and energy applications, for example geothermal energy, carbon sequestration, and waste disposal, require a deeper understanding of geomechanical properties, mineral transformations, and fluid-rock interactions. How does the geologic evolution of shallow crustal conditions (sediment, rock, and fluid) influence the system response to anthropogenic forcing associated with energy production, CO2 storage, and waste disposal and how do these factors impact induced seismicity and earthquake hazard?

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

   - Our communities lack the databases on Earth materials required to address key problems in mineral physics & rock deformation. Ready access to such data will facilitate transformational interaction across fields and promote innovative assessment of key problems. We need to establish links between future COMPRES and DEFORM databases and those on seismic, petrologic, thermodynamic, elastic, geochemical, and crystallographic properties of Earth materials.
   - Lack of reliable and sufficiently-automated data analysis software able to push the limits of quantitative information retrieval from both experimental and theoretical data sets to new limits of spatial, temporal, stress resolution and system complexity in response to revolutionary improvements in experimental technology capabilities. Data volumes and real-time signal processing are currently growing far faster than post-processing techniques, and automated methods for data analysis are needed to meet this challenge.
   - Need for workflows, data mining capabilities, intuitive data systems, and easy to use web-based tools that encourage best-practices in reproducible science and transparency in data processing and storage.
   - A key roadblock is that of how to extract scientifically useful information on rock fabric and mineral textures from images collected using a variety of methods. We need to build on recent developments in microtomography and 4D imaging
   - Need for accessible, easy to use computational tools that would enable calculation of physical, structural, thermodynamic and transport properties of Earth materials at any pressure and temperature conditions, particularly those not easily accessible through experiments.

TECHNICAL INFORMATION/ISSUES/CHALLENGES

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

   - Central data system for DEFORM and COMPRES science. This should include storage, visualization, and search protocols to provide community access to our data and solutions that will reduce activation energy to including data in these databases.
   - Community technical forums, including websites, focused on CI developments for both DEFORM and COMPRES. We note that COMPRES has a Technology Office at Argonne National Laboratory and a technology-oriented website maintained by COMPTECH, which could be a starting point. We need both tools to compare data from different labs, including functional fits, statistical analysis and model evaluation and a social network associated with our data system to provide a forum to interact virtually and lower barriers to interdisciplinary interaction between researchers. These tools should include a way to capture information about how users interact with the databases, and automated methods to improve the data system based on this information.
   - Central archive of experimental samples with integrated workflows, database templates, and community-wide DOI system for samples
   - Automated system for storage and evaluation of microstructural images, including rock fabric, texture evaluation and pore networks, as well as comparison of laboratory and field microstructures and shear zone texture.
   - Extending data mining capabilities/tools and interlinking existing repositories, (e.g. crystal structure and spectroscopic databases) with newly developed databases.
   - Reliable, sufficiently automated, easily accessible and well-documented software for efficient (preferably real time) processing of large volumes of experimental data and results from theoretical and numerical studies.
   - Improve accessibility of high-performance computing (HPC) by both lowering the entrance barrier and providing analytic/query tools to make the results of these calculations readily available to the wider observational and experimental Earth science communities. Collaboration/assistance from HPC staff with earth-science researchers at HPC centers.
   - Create a comprehensive reference Earth model that includes both deformation and elastic properties.

COMMUNITY NEXT STEPS

1. **List of what your community needs to do next to move forward how it can use EarthCube to achieve those goals:**

   - Develop RCN's for DEFORM and COMPRES databases, data systems, and data sharing among data sources. Develop a database prototype with functionality to link databases and record data processing with the goal of moving toward a system for reproducible science. Develop protocols for data transfer between data sources (experimental and computational) and data consumers.
   - RCN proposal for designing next generation data collection and analysis tools. Explore visualization and analysis tools that are known to other communities to see if they can be applied to our problems. An important component of this is image analysis and pattern recognition

including automated methods to identify deformation microstructures, LPO and other examples of fabric in rock and minerals.

- Utilize EC Building Blocks and other available funding mechanisms to enhance and expand the most promising existing CI solutions useful for our communities.
- Engage discussions with representatives of existing community organizations and facilities (DEFORM, IEDA, COMPRES, HPCAT, GSECARS, ANL, ALS) to take advantage of existing resources (websites, videoconferencing capabilities, technology focused personnel).
- RCN between experimentalists, seismologists, geodynamicists, geochemists, and computer infrastructure experts to develop approaches for achieving a three-dimensional reference Earth model that provides a mechanism to link geophysical observations with laboratory and theoretical studies.

Solar-Terrestrial Research

Executive Summary:

Science-Driven Cyberinfrastructure Needs in Solar-Terrestrial Research

Held at New Jersey Institute of Technology, 2014 Aug. 13-15

Steering Committee: ***Gelu M. Nita, Dale E. Gary, Andrew J. Gerrard, Gregory D. Fleishman, Alexander G. Kosovichev, Vincent Oria, Marek Rusinkiewicz***

Introduction

More than 80 domain scientists and students from three sub-disciplines of Geospace research (solar/heliospheric, magnetospheric, and upper-atmospheric research), as well as computer science, met at the *Center for Solar-Terrestrial Research* at *New Jersey Institute of Technology* for a 3-day workshop to examine the field's current state of cyberinfrastructure (CI) and its future needs. To prepare for the workshop, the steering committee identified 17 CI-knowledgeable leaders (listed at http://workspace.earthcube.org/solar-terrestrial-end-user-workshop/) who represent each of the NSF Geosciences programs SHINE, GEM and CEDAR, as well as computer science. This scientific organizing committee identified an additional 40+ scientists for invitation to the workshop, as well as NSF program managers Eva Zanzerkia (Earthcube), Ilia Roussev (SHINE), Anne-Marie Schmoltner (CEDAR), and Raymond Walker (GEM).

We endeavored to balance the demographics among the sub-disciplines and in relative experience of the participants. Approximately 25% of participants were early-career (8 students, 7 young scientists), 25% mid-career, and 50% in senior positions. The sub-discipline participation was nearly evenly split, with 34% SHINE, 23% GEM, 23% CEDAR, and 19% computer science. The preponderance of solar participation reflects mainly the concentration of solar research among the local NJIT participants. The organizers believe that the workshop successfully captured the expertise and experience of the Geospace research community, and that the findings herein represent the consensus view of leaders and practitioners in science-driven cyberinfrastructure among space-science researchers.

The Geospace disciplines are somewhat unique in the Geosciences for at least two reasons: (1) the disciplines are dominated by highly dynamic phenomena, and hence the data are organized mainly (though not entirely) on events and time rather than primarily spatially; and (2) the science drivers in these disciplines are studied in depth and decided upon as a broad-based community endeavor culminating in a decadal survey report every 10 years. The most recent report, *Solar and Space Physics: A Science for a Technological Society* (National Research Council, The National Academies Press) was released in 2013, and serves as the main guide for science drivers examined during the workshop. None of the findings below are meant to conflict in any way with the national science goals outlined in this decadal survey.

In addition to science goals, the NRC Decadal Survey also recommended, as a high priority, the implementation of an integrated initiative (DRIVE) to develop critical new technological capabilities in order to address the decadal survey's complex scientific topics. In particular the decadal survey encourages the development of a "data environment that draws together new and archived satellite and ground-based solar and space physics data sets and computational results from the research and operations communities." This included "community oversight of emerging, integrated data systems" and "exploitation of emerging information technologies" with "virtual observatories as a specific component of the solar and space physics research-supporting infrastructure."

Science Issues and Challenges

Important science drivers:

The latest NRC Decadal Survey in Solar and Space Physics outlines four overarching key science goals for solar-terrestrial studies in the coming years. Below are more-focused science goals, consistent with the Decadal Survey goals, that we anticipate will benefit most from investments in cyberinfrastructure during the next 5 - 15 years:

- **Understanding the couplings among physically different domains ranging from the solar interior to the Earth's atmosphere**: The advent of "Big Data" (the aggregation of large, complex, heterogeneous data sets) in observations and numerical modeling holds promise for rapid progress in solar-terrestrial research. Space- and ground-based observatories will provide important constraints for models in terms of boundary conditions and synthetic observables. New observational data and computational advances provide new opportunities to develop cutting edge, data-driven models for the evolution of the magnetic flux below and above the solar surface, its influence throughout the heliosphere, and its impact at Earth. New cyberinfrastructure is required to improve our knowledge of the transfer of physical drivers across different physical domains from observational data and numerical simulations.

– **The study of the fundamental processes through which magnetic energy is generated, stored, released, and propagated**: This is critically dependent on an advanced cyberinfrastructure that enhances our ability to assemble, analyze, and visualize multi-instrument, multi-wavelength datasets covering multiple temporal and spatial scales in combination with detailed physical models. The application of computer vision and machine learning techniques to identify features across different physical dimensions and to better mine large, distributed databases will be needed to enable event identification and statistically driven analysis. Of particular interest is understanding the process of magnetic reconnection, the primary mechanism for energy release in solar flares and coronal mass ejections, which controls the occurrence and severity of magnetic storms through transport of mass, energy and momentum both at the sunward side of the magnetosphere and in the magnetotail.

– **Predicting the solar wind and Interplanetary Magnetic Field in the near-Earth environment.** Understanding the origin of magnetic flux structure at the Sun, and how it evolves during magnetic eruption and propagation through the heliosphere to produce the relevant spatial scale of B_z variation near Earth that drives magnetic storms, will depend critically on in situ and remote sensing observations from the *Solar Dynamics Observatory*, *Magnetospheric Multiscale, Solar Probe Plus* and *Solar Orbiter* and other spacecraft, as well as ground-based facilities, combined with modeling techniques capable of simulating CME flux ropes from the Sun to the Earth. The many disparate types of data and the broad range of spatial and temporal scales involved in both observations and models present a substantial cyberinfrastructure challenge.

– **Understanding the acceleration of particles throughout the Sun-Earth system.** Acceleration of electrons and ions, often to extremely high energies, is ubiquitous throughout the solar atmosphere, heliosphere, magnetosphere, and ionosphere, and creates hazards for humans and technological systems (spacecraft, communication and navigation systems, and even aircraft) everywhere within Geospace. In every region, important tasks remain, such as: identifying the acceleration mechanisms that operate in the various regions of the Sun-Earth system; determining which mechanisms are most important at different times and locations; identifying common *vs*. distinct mechanisms in different regions; identifying the more important plasma instabilities that operate in the different regions and the role they play in particle acceleration under varying conditions; and following the propagation of accelerated particles within and across regions of the Sun-Earth system.

- **Understanding and forecasting the effects of forcing** on the coupled Ionosphere-Thermosphere-Mesosphere (ITM) system. The ITM system presents a unique challenge in that strong coupling between charged and neutral species dominates physical processes. The system is responsive to external forces, e.g. reconnection, which impose global electric fields and magnetic currents, but also to internal processes, e.g. tropospheric heating and upward transmission of tidal forces, ionospheric instabilities, ion-neutral collisions and frictional drag. The coupled system demands cross-disciplinary study involving data acquired over multiple time and distance scales from ground and space observatories. Our ability to facilitate telecommunication and navigation, prevent catastrophic failure of the power grid during magnetic storms, or protect space assets from collisions demands accurate forecasting of the ITM response to forcing. Unique to this effort, international collaborations often require the participation of poorer countries with desirable locations for observations, but without the means to install instrumentation or distribute data in optimal ways.

Current Challenges to High-Impact, Interdisciplinary Science:

The main challenges identified by workshop participants center around bridging the gaps among the Geospace sub-disciplines, to foster interdisciplinary research.

Challenges in finding / discovering data

- Users do not know how to search for data across multiple repositories, and in general what data sets/resources exist. Data are hard to find, and even harder to transform into the form needed for further analysis.
- Semantic techniques should be available to enable broad discovery and use of data. Tools/libraries that enable the generation of metadata (annotations) in an automated fashion would be preferred.
- Joint data discovery ideally makes use of of centralized data repositories or search facilities where all the metadata (and pointers to the data) are queried and made available through a common interface. Complementary to this would be the implementation of a system based on semantic web technologies, which would require that a widely accepted standard vocabulary/ontology (suitable for our community) be put in place that the community agrees to abide to.
- There is a need for encouraging adoption and consistent usage of metadata standards for the essential attributes of both observational and modeling data sets, as well as an agreement on vocabulary to use.
- Getting to a set of “widely accepted standards” is itself a challenge. Also needed are translation tools (“ontology alignment”) between different sets of standards, especially where there are already multiple sets of established practices.
- The Geospace disciplines increasingly need better tools for mining our spatiotemporal datasets for features, both known and unknown
- The tools need to be scalable, to work for both large and small datasets.
- Data query: enabling the easy and effective querying of very specific subsets of data in order to tailor the results according to a specific science objective, thus reducing the volume of the data transfer. Good metadata and strong quick-look tools play a big role in this.
- Data volumes are becoming prohibitively large. It is not feasible to co-locate all data sets, or even apply the “old model“ of requiring users to download all the datasets of interest onto their own computers to manipulate them locally. Analysis increasingly needs to be co-located with the data, but this is problematic for analysis of multiple datasets, located in different places. Processing and user-driven analysis carried out at these large data centers may provide a solution to this coming problem, but mechanisms need to be in place to allow these providers to develop and support these (potentially costly) capabilities.

Challenges in working with data

- Continuity of data sets (both space and ground-based) over time has an increasing value as our ability to mine and probe these large data collections grows. Ensuring continuity should be a factor in funding decisions. (For example, there are concerns about several older instruments with no successor at the moment.)

- There are similar issues of continuity in the development of data analysis tools as well as instruments.
- Getting the most out of existing or legacy data; ensuring things do not get lost over time as missions or groups end.
- Information about assumptions, sources of error, and methodologies should be included along with the data.
- Need methods to ensure scientific reproducibility by allowing citation of specific data products and processing steps used in a scientific study.
- Need a mechanism for ensuring proper attribution of data sources in publications. It is critical to record provenance of all data to improve future reuse.
- Need better benchmarking/validation of data catalogs for researchers in different disciplines: it is important to have clear quality metrics that allow users to determine which data points are "good" or "bad" for their purposes.
- It is important not to "re-invent the wheel." If someone has "solved" a problem, other communities need to be able to find out about this and make use of it.
- The wide variety of analysis tools and languages in current use inhibits the development of a common set of analysis tools. Clearer documentation and use of software development best practices would help mitigate this confusion.
- There is a need for a strong leadership structure: a project should be run by a single, strong entity with broad community buy-in to ensure coordination.

Challenges in cross-disciplinary science / working with data outside our sub-discipline.

- Data from outside a researcher's field is difficult to find and learn how to analyze.
- An impediment to cross-disciplinary research is that while the same problems might be studied in different sub-disciplines, the observables, scales, and parameter regimes may be quite different.
- It is difficult to find sources of funding for cross-disciplinary research.
- Researchers using data from outside their areas of expertise need trusted catalogs of events and categorizations
- Data integration is needed to enable interfacing and interoperability among diverse datasets.
- Need better support for 'sun-to-mud' efforts. Solutions may be to have more common workshops, and classes offered online by multiple institutions.

Modeling-specific challenges

- It is important to compare and address discrepancies between data and models. Tools are generally not readily available to directly compare model outputs and observations.
- If these tools were available, iteration between modeling and data comparison could take place, allowing ongoing improvement of both.
- While data are often open and analysis code is sometimes open source, the same is not generally true for models (although it should be).
- In terms of modeling: there is a need for better flexibility/modularity in large model design so various groups could "plug and play" their components.

Educational, societal, and public outreach challenges

- There is a dearth of data-science and cyberinfrastructure-related content in the domain-specific academic curricula, impairing the ability of students to incorporate existing tools and best practices into their research.
- Scientists often do not know how to scale up their cyberinfrastructure usage from the desktop to make use of high-performance computing (HPC).
- Students and practicing researchers need training on how to use GPUs and other advanced computing resources.
- Scientists want to share their data in the public domain, but may worry about potential misuse or misinterpretation of the data.

Technical Issues/Challenges

Many of the interdisciplinary science challenges noted above are rooted in technical issues that must be addressed in order to successfully overcome them. The breakout sessions devoted to technical challenges included moderators who are computer scientists, in order to encourage new thinking.

- There is a need to develop computationally efficient capabilities for searching and expressive querying of Large/Diverse/Distributed Data Sets including provenance and data quality. What is of interest to scientists can be very complex to define. With today's high-volume databases, it is increasingly important to locate and download only the portion of data of interest. Propagation delays from one regime to another within the Geospace system make event searches challenging—e.g. how to do correlations to find linked events among data sets with such delays, without downloading all of the data.
- There will be a continuing need to discover, search, and utilize historical datasets, which must be preserved and, if necessary, modernized through metadata indexing to bring them into discoverable form.
- Data providers, especially new and actively maintained services, need to include well-documented APIs (application programming interfaces) and service interfaces, to aid in development of flexible workflows for utilizing the data resources.
- Some metadata standards already exist, but translators/converters are needed for searches bridging solar-terrestrial environments (solar, heliosphere, magnetosphere, ionosphere/upper-atmosphere) to promote interdisciplinary science. Additional efforts to agree on a wider standard of keywords, vocabulary and ontologies would be useful, but difficult.
- A platform and standards for data and software citations need to be further developed and widely adopted. A scheme for searching ranked databases and software according to popularity, usage, and quality would be a useful addition.
- Workshops/tutorials and academic curricula are needed to teach standard tools and techniques for interdisciplinary research to the community (e.g., orbital discovery tool). Community-developed toolkits (e.g. those at SolarSoft, sunpy.org, itk.org) are important sources of cross-platform tools for general analysis. Community involvement in further open-source tool development (e.g. through Github) should be strengthened and encouraged.

- Tools are needed for generalized Event/Object recognition in space and time, and for visualizing multi-dimensional data in large data volumes

Community Next Steps

Since this Solar-Terrestrial Cyberinfrastructure workshop occurred rather late in the process of Earthcube governance, we have the advantage of knowing the context of the program within which we should coordinate our efforts. Many of the challenges identified during the workshop have also been identified by other domain workshops, and hence our community can form Earthcube working groups or join with others already forming within Earthcube. In addition, our community can undertake the following steps, and also encourage NSF to provide Earthcube funding opportunities to address these areas:

Tools and Standards

- Make/collect a list of useful tools and services (with user reviews)
- Provide additional tools for generating metadata from existing data and manipulating metadata in the form of plots, indexing
- Support development of community-led general analysis toolkits
- Provide translators between standard data formats
- Provide translators between metadata (e.g. keywords) standards
- Develop standard service interfaces (such as APIs)
- Develop “one-stop-shopping facility” to aggregate data, or facilitate ordering/delivery of data

Cross-disciplinary CS/domain scientist collaborations

- Assemble domain scientists and computer scientists to attack specific and realistically achievable high-value science goals as identified by the decadal survey
- Identify and list the most widely-used data-sets in the relevant disciplines and design data integration tools according to the above-mentioned science goals
- Create hyper-dimensional visualization tools
- Develop the capability for advanced semantic queries for nearest-neighbor matching of widely dissimilar data
- Develop the capability to construct queries of what is missing (identifying gaps and dealing with intermittency in data coverage)

Education (community and academic)

- Adding cyberinfrastructure and computer visualization components to solar-terrestrial curricula.
- Educating domain scientists on scaling up their applications from desktop to HPC
- Access to HPC resources for training in solar-terrestrial research
- Education on how to utilize GPU and other advanced computing resources
- Advanced data analysis techniques (e.g. inverse theory, forward fitting, data assimilation)

Data management

- Searching and querying long-term archived databases with access control and provenance
- Use of DOIs and alternatives for data and software citations
- Tools and standards for creation of metadata that tracks database use (who, for what purpose, popularity)
- Cloud storage and HPC processing
- Support for creation, population, and operation of new databases based on new instruments and modeling efforts
- Capability for creation of quick-look data products

Model input/output

- Develop techniques for data-assimilation, data-driven modeling, and cross-domain model coupling
- Metadata concepts for model output (descriptive of format)
- Develop standards and guidelines for making model output shareable and comparable
- Search tools for integrating observational and model output data

Quantifying data quality

- Include valid error estimates together with data
- Include information about data quality, completeness, and fitness for use
- Research methods and practices for quantifying errors (random, systematic)
- Biases introduced by data processing

Encouraging good practices

- Study feasibility of creating cloud-storage for data, whose use would enforce good practices as a prerequisite for use
- Create or join an Earthcube working group to identify and share information and tools for enforcing metadata standards
- Include software engineering and development techniques as part of academic training

EXECUTIVE SUMMARY: EARTHCUBE SEDIMENTARY GEOLOGY WORKSHOP

Marjorie Chan - University of Utah, and David A. Budd - University of Colorado, co-conveners;
March 25-26, 2013, University of Utah

Earth Cube Workshop Title: End-User Workshop For Sedimentary Geology

INTRODUCTION

The Sedimentary Geology Community (SGC) domain workshop brought together 57 geoscientists with expertise in modern and ancient sedimentology, stratigraphy, basin analysis, paleontology, paleoclimatology, sedimentary geochemistry, sedimentary petrology, petroleum geology, paleopedology, and geochronology. This community has historically focused on research questions related to the processes that form, shape and affect the Earth's sedimentary crust and distribute key resources such as hydrocarbons, coal, and water. Sedimentary geoscientists also use the sedimentary record to explore the continental crust's evolution, the dynamics of Earth's past climates and oceans, and the evolution of the biosphere. Sedimentary systems also form the framework for the research conducted in many other geoscience communities.

Prior to the workshop, participants provided statements on overarching science drivers in the field, challenges to integrating the community into EarthCube, and the research themes that could be pursued with an ideal EarthCube. Breakout sessions were held on scientific drivers, impediments to sharing and using data, current cyber resources, needed data sets and tools for the future, and potential impact of EarthCube on SGC teaching.

SCIENCE ISSUES AND CHALLENGES

A. **Important science drivers -** Three overarching societal issues were highlighted as drivers that will condition research within the SGC community over the next 5-15 years. Multiple scientific challenges were identified relative to each driver. The primary theme of SGC is to fully integrate our discipline with Earth, Atmospheric, Oceanic, Biologic, and quantitative sciences in addressing the sedimentary dynamics of Earth and planetary systems from the beginning of time, and the current role of human interactions into a sustainable future.

Driver #1 - Securing the energy and water resources needed for an increasing global population while balancing resources for a sustainable Earth.

Related research challenges within the SGC community:

1. Predicting lateral spatial heterogeneity in the geometry and physical properties of sedimentary rocks/bodies. This is a necessity for effectively predicting resource distribution, modeling fluid flow, and mitigating contaminant problems.
2. Improved understanding of organic-rich fine-grained sedimentary systems throughout geologic time, particularly their origin and the processes that generate sedimentologic and geochemical heterogeneity.

Driver #2 - Understanding the Earth as a system, the nature of global climate change, and its impact of climate change on life, the environment, and Earth resources.

Related research challenges within the SGC community:

3. The deep-time sedimentary record must be scrutinized to learn how the Earth's climate system operates in periods of stasis, rapid change, and greenhouse and icehouse conditions. This requires:
    a. Continued development of proxies for ancient climate, improvement in existing proxies, and reconciliation between proxies.
    b. Analyzing the sedimentary record to identify and understand the components of deep-time climate change (forcing factors, feedbacks, tipping points) and the resultant impact on the deep-time Earth system (changes to hydrologic cycle, weathering, denudation, sediment fluxes, nutrient runoff, ocean circulation, extinctions and originations of life, etc.).
4. Develop a deeper understanding of the interplay between life, the physical and chemical environment in Earth's past, climate, tectonics, environmental change, and sedimentary processes.

Driver #3 – Human activities influence, and in some cases dominate, many Earth surface processes. Understanding those anthropogenic influences will be necessary to minimize risks to society and insure environmental sustainability, particularly in deltaic, coastal zone, reefs, lake, and fluvial settings.

Related research challenges within the SGC community:

5. Development of morphodynamic models of how sedimentary environments and landscapes responded on daily to millennial scales to climate change, sea-level rise, sediment supply, induced subsidence, engineered structures, etc.
6. Determining how to use the sedimentary record to make predictions about future environmental changes, assess critical boundary conditions, quantifying parameters of environmental change, and evaluate rates of change.

Research Challenges Common to Drivers #1-3 above:

7. What controls stratigraphic architecture and landscape dynamics? Revisiting the respective roles at varying temporal and spatial scales of autogenic (intrinsic feedback loops) vs. allogenic (climate, tectonics, eustacy) controls.
8. Development of geochronological tools that provide more precise and accurate timing of critical events in Earth's history are necessary to meet all other research challenges. Geochronology must address: (i) the timing, duration, and rates of ancient climate change; (ii) how rapidly life responds to environmental change; (iii) the rapidity of geochemical changes in the Earth system; and (iv) the recurrence & magnitude of natural hazard events preserved in the sedimentary record.

b. Current challenges to high-impact, interdisciplinary science - The workshop specifically focused on challenges and impediments to sharing and using data and other cyber resources. Results are loosely grouped with respect to data bases, cyber structure, and people issues. ***Those that maybe unique to this community are in bold italics.***

<u>Data</u>

1. Lack of knowledge regarding what database and tools exists.
2. Access to data, particularly subsurface data in the private sector and legacy data in physical collections, theses and dissertations, and gray literature, etc.

3. ***Inadequate documentation of data (lack of location data, meaning of symbols on graphical data, unstated uncertainty and reproducibility, no stratigraphic/facies context, incomplete age information, is it raw or corrected data, how was it corrected).***
4. Coordinate systems in metadata are not uniform.
5. Lack of specified methodologies (unified data paradigm) – are different data sets really comparable?
6. Concerns about quality and authenticity of data, particularly older data.
7. ***Uncertainty in whether errors have been removed and data updated (changes applicable to dates, stratigraphic nomenclature, taxonomy).***
8. Data discovery across organizations is difficult (impossible).
9. Coupling diverse data sets is hard.
10. ***Uncertainty in observational data like measured stratigraphic sections – measuring scale (resolution) typically unstated, unclear if lack of features is due to absence or failure to record; what is recorded is subject to interpretation and expertise; definitions/classification of features may vary; no consistent format for representing features/data.***
11. Inability to search/query both observational and interpretive data.

Cyber structures – they do not mimic science workflows and thought processes

12. Lack of catalogues as to what is held by organizations; no portal that provides all the need connections.
13. Lack of interoperability between data sources due to vendor's proprietary data formats.
14. Steep learning curves to use cyber resources (not user friendly due to overly specialized formatting, processing, unintuitive interfaces, difficulties in uploading, etc.).
15. Lack of uniform formatting and standard; too rigid a data entry form to capture what is needed.
16. Existing resources not easily searchable, especially for information by place/area, time interval, type of object (e.g., type of facies, environment, sedimentary structure).
17. ***No easy way to integrating subsurface and surface data.***

People

18. Lack of incentives to data share.
19. Reluctance to share pre-publication of data, interpretations, and implications (cannot get a citable DOI for just a data set).
20. Lack of training needed to use cyber resources.
21. Lack of time and resources to maintain a website, data base, or tool.
22. ***Concern about unethical uses of one's data (e.g., GPS coordinates of fossil and mineral localities makes poaching and theft for commercial purposes possible).***

TECHNICAL INFORMATION/ISSUES/CHALLENGES

A. Existing tools, databases, etc. needed for pursuing key science questions - The Paleogeoscience domain workshop previously identified ~140 cyber databases, repositories, or tools, with particular focus on paleobiology, marine sediments, geochronology, and paleoclimate. The Sedimentary Geology workshop added 83 additional cyber resources to the compilation – 38 databases, 17 repositories, and 28 tools. These additions particularly focused on LiDAR, map resources, and tools for use in sedimentary geology and subsurface analysis. Of particular note is that there are very few databases for onshore sedimentary geology, most repositories of subsurface data are state or federal agencies, and the most thorough software tools are commercial.

B. Desired tools, databases, etc. needed for pursuing key science questions - To forge new ground and develop richer comprehensions of complex problems and systems, sedimentary geology research requires multidisciplinary approaches, easy access to large volumes of geologic and geophysical data, better integration of that data and legacy data, and increasingly sophisticated numerical modeling of sedimentary systems and stratigraphic architecture.

A Google Earth-like interface is envisioned with topography, surface, and subsurface geology. The interface would (i) allow a wide range of queries, (ii) compile and visualize a variety of data for different time intervals and geographic locations, and (iii) have the ability to create cross sections from designated line paths and make maps for designated areas and time/depth intervals.

Databases[1] *(Geo-referenced; can also be catalog information, not just data)*

1. Geologic maps, cross sections, seismic and GPR lines, LiDAR data, macrostratigraphy.
2. Distribution of fossil organisms through space and time (e.g., Paleobiology Database).
3. A better compilation and integration of the available paleoclimatic data.
4. Drill hole that integrates or links across state boundaries and includes locations, formation tops, geophysical logs, cored intervals, core photographs, poro-perm data, thin section imagery, total organic carbon values, thermal data (e.g., vitrinite reflectance).
5. Measured sections of outcrop and core (both referenced by midpoint of section or line in dipping units). Include scanned images of cores, lithologies, sedimentary structures, grain sizes, textures, fabrics, contacts, trace fossils, thin section imagery, poro-perm data, mineralogy and whole-rock geochemical data (e.g., stable isotopes).
6. Sedimentary rock imagery: stratal geometries, sedimentary structures, photomicrographs, etc.
7. Data on age constraints of stratigraphic units, including source and basis of age.
8. Hub for coordinating databases.

[1] The SGC recognizes the need to develop protocols for metadata, as well as protocols and formats for core and measured section databases.

Search Capabilities

9. Multi-tiered search engines to access and search different databases.
10. Searchable map-based areas of interest by time, space, stratigraphic unit or topic.
11. Spatial querying for published work.
12. Filtering tools for searching (search engine and tagger).
13. Ability to search by example - an image of the object or a verbal description - and the query system finds things that are similar (fuzzy query for dark data?).

Tools *(must enable range of data formats and conversions)*

14. Template or checklist tool for metadata format.
15. A suite of tools to easily sort/analyze data using available metadata.
16. Ability to map (with contours) all types of quantitative data.
17. A set of tools that will correlate between sections/core.
18. Tools for compilation and correlation of biostratigraphic ranges for different index taxa.
19. Basic sedimentary interpretation tools (e.g., of depositional environment) involving guided questions that direct interpretation process.
20. Capability of determining sediment volumes/thickness/accumulation rates/fluxes from measured sections, logs, seismic data, etc.
21. Open source visualization software for well, seismic, and LiDAR data.
22. Open source visualization software for stratigraphic columns, timescales and other data (biostratigraphic, chronological, geochemical, petrographic, etc.).
23. Higher resolution paleoclimate climate models.
24. Interoperability with the CSDMS (Community Surface Dynamics Modeling System) suite of modeling tools.

25. Ability to track users of particular features to help organize conferences and workshops of people with common interests.

Other

26. Ability to enter the data as it is collected.
27. Continued development of GeoDeepDive techniques - machine learning data-mining to extract info from PDFs and convert it to a database that can be directly queried.
28. Training modules for database creation and entry, search tools, analysis and visualization.

C. Potential impacts on education and workforce preparation:

The SGC community recognizes that EarthCube can revolutionize how we teach by focusing on research themes and rich real-world data. But educational applications will need to have an ease of use and shallow learning curve so that student can use the data and tools with limited training. EarthCube can be a teaching tool that will supports project-based courses, help build critical thinking skills, promote student inquiry at all levels, and demonstrate how to do research. Participants saw great potential in using EarthCube to explore the spatial and temporal scales of investigation unique to the Earth Sciences. For the non-STEM student, it can change the way they think about how science is done.

Two proposed examples of how the SGC community might use EarthCube in the classroom:

- In a lecture topic on sand size in rivers, students could use the GIS-module and zoom to multiple localities. They access information on the grain sizes and average month discharge (proxy for average flow velocity) at various points in numerous rivers. Students then explore patterns, distance from the source, relation to flow velocity, slope of the land surface, etc. in order to assess controls on grain size (do they decrease with distance from the source as the lecture asserted?).
- In an assignment to analyze the paleogeography in the Cretaceous of the western U.S., students choose sites (GIS module) and retrieve stratigraphy, biostratigraphy, detailed measured sections that show lithology, grain size, sedimentary structures, biofacies, etc. They then use a cross-section building tool and a paleogeographic mapping tool to make their own reconstructions through time.

There was also strong enthusiasm for EarthCube to be a vehicle for virtual field trips via 3D rendering of outcrops and fly-over tools. Field geology requires students to think in four dimensions, evaluate multiple working hypotheses, work at multiple spatial scales, and draw conclusions from incomplete data sets. Virtual trips in EarthCube have the potential to achieve the same results, but also be more comprehensive in that the field component can be supplemented by ready access to appropriate analytical data, thus making the "trip" a comprehensive analysis and investigation. Such opportunities make it possible for all student geoscientists, including limited mobility students and non-traditional students, to visit and "work" any location.

COMMUNITY NEXT STEPS

List of what your community needs to do next to move forward how it can use EarthCube to achieve those goals:

- In order to improve communication with all members of the community, a listserve was immediately established after the workshop (SedGeoNet listserve). It is anticipated that this listserve will be turned over to the new STEEPE coordinating office (*Sedimentary Geology, Time, Environment, Paleontology, Paleoclimatology, and Energy,* http://www.steppe.org/).

- The STEEPE coordinating office will be encouraged to develop structures (web-based dialogs, workshops, eNewsletters, virtual "idea" fairs) that will move the community forward on EarthCube related activities

- Four potential Research Coordination Networks (RCN) initiatives and one Building Block initiative were identified for development (see below). Each would include geoscientists from other domains (particularly structure, tectonics, igneous/metamorphic petrology, geochemistry, hydrogeosciences, geophysics, geomorphology).

RCN Initiative - Geoscience images

An RCN focused on imagery will bring researchers and educators together to: identify the range of scientific images (fields of view that span kilometers to nanometers), converge on common understanding as to what information should be extractable for images in EarthCube, and begin a dialog on how to create interoperable image databases that can be searched and exploited.

RCN Initiative - Framework and user interface for collecting field data

Field data is typically viewed as point data (objects in space and time), but for sedimentary geologist the measured section can also be viewed as streaming data – multiple tracks of different data collected progressively up the outcrop. The challenge is capturing this data digitally in real time so that in can be retrieved and used for varied applications. This issue needs to be pursued in coordination with other field-based geoscience domains and must integrate with cyberinfrastructure experts from the start so the field workflow can be captured and integrated.

RCN Initiative – Subsurface data integration

Tremendous amounts of subsurface data already exist. These data are very diverse. Some is streaming, some is point data. It includes geophysical and geochemical data; images, actual samples/cores, analytical compilations, and derivatives (maps, cross sections, etc.). Some are digital others are analog. Much is already is disparate databases (particularly those of state & federal agencies), but none are interoperable. Georeferencing is varied. Data discover and access is random.

The goal of this RCN would be to what data needs to be made more accessible, how it might be linked, and how it needs to be visualized. The RCN would have to include cyberinfrastructure experts so as to make the connection between how the data can be managed versus how geoscientists want to access and use this data.

RCN Initiative - 3D Geodata and its visualization

There are many groups across geosciences that are interested in 3D visualization of surface and subsurface data. An RCN amongst these groups would focus on issues surrounding spatial resolution, temporal resolution, metadata standards, and the development of tool that better integrate surface data (2D maps) to subsurface data. Exploitation of existing tools/databases would be included. A goal would be to assess issues such as what types of data, what types of visualizations, and workflows. The long-range target is a portal with multiple visualization tools, demonstrations on each tools capabilities and applications, and training modules that make each tool broadly accessible.

Building Block Initiative – Mini EarthCube - Mesozoic Geology of the Colorado Plateau

A "mini EarthCube" project will integrate current data sets to build an architectural, geospatial visualization model (a location-based "Google Earth"-style search engine) of the Mesozoic sedimentary geology on the Colorado Plateau (Utah, Colorado, Arizona, New Mexico). This will be a proof of concept resulting in a geoinformation framework with a portal that allows access and visualization of fused/ tiered/multi-scaled geological layers. It will serve as a test case for the future, grander EarthCube. This "mini EarthCube" initiative requires partnership among various entities within and beyond the SGC, and integration with GIS cyberinfrastructure experts. We propose building a model patterned after the successful, existing Lunar Mapping and Modeling Project (LMMP) – www.lmmp.nasa.gov, in collaboration with the cyberinfrastructure team that designed and delivered the lunar model to NASA for use by the lunar science community, educators, and the general public. This collaborative approach will leverage on NASA and NSF-funded capabilities and approaches that can be the springboard to quickly move to a focused geologic project that will have high impact.

EXECUTIVE SUMMARY: EARTHCUBE WORKSHOP RESULTS

Mohan Ramamurthy, Unidata/UCAR, Fuqing Zhang, Penn State U., and Russ Schumacher, Colorado State U.

17-18 December 2012

Earth Cube Workshop Title: Shaping the Development of EarthCube to Enable Advances in Data Assimilation and Ensemble Prediction

Introduction (field(s)/area(s) of interest and purpose, number of participants)**:** Data intensive science has rapidly emerged as the Fourth Paradigm of scientific discovery after empirical, theoretical, and computational methods. This is particularly true in the area of data assimilation and ensemble prediction. This workshop was held to shape the development of EarthCube from the perspectives of the mesoscale modeling, data assimilation, and ensemble prediction communities and help in building an infrastructure that makes it easy to integrate and use observations and model output from disparate sources, support distributed modeling and data assimilation activities and share the resulting data, allow investigators to perceive linkages that today are obscured because data formats are incompatible, increase data transparency and ease-of-use, and reduce "time to science and publications."

There were 72 registered participants, and they came from all sectors of the atmospheric science community (academia, government, and private sector), and from geographically distributed universities, research labs, and organizations that provide data to the atmospheric research community. Since the workshop was held in Boulder, CO, where NCAR and NOAA/ESRL labs are located, the workshop benefited from a large number of local participants.

SCIENCE ISSUES AND CHALLENGES

1. **Important science drivers and challenges:** Participants identified several high-priority science questions that will be the focus of interdisciplinary efforts during the next 5-15 years (list 3 to 6).

   - What are the limits of predictability in the atmosphere? What are the sources of uncertainty/ errors, and how do they feed into predictability?
   - What observations are critically needed to enhance atmospheric predictions, and where? What is the optimal configuration of the observation network?
   - What are the appropriate types, combinations, and configurations of parameterization schemes for high-resolution mesoscale models? How can the errors and biases in these parameterizations be quantified and corrected?
   - What is the optimal ensemble configuration to accurately predict the distribution of possible outcomes? How many ensemble members are needed and how should the ensembles be initialized?
   - What are the advantages and disadvantage of variational versus ensemble-based data assimilation techniques, as well as different types of hybrid approaches?
   - What are the most effective ways to post-process ensemble forecasts to achieve reliable and calibrated probabilistic predictions?

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).

- Significant barriers exist in using the data efficiently or integrating them into data assimilation or ensemble prediction systems. Today, there is too much overhead to doing research efficiently – e.g., setting up one's data and analyzing it. Rarely are there tools that really reduce this overhead.

- The scientific community lacks easy-to-use common cyberinfrastructure frameworks, data format standards, sufficient metadata for observations, and methods/tools for quality controlling observations, mining of large volumes of data, visualization, and verification.

- While many good facilities exist in this field (e.g., Unidata, DTC, and DART), they sometimes operate in silos and their activities and services are not always well coordinated or integrated.

- Lack of a central repository for finding, accessing, and using data and software.

- Significant spin-up time for students in preparing, using, processing, and analyzing data. While similar challenges exist for researchers, such problems are particularly acute for students who have a limited time before they graduate.

.

- Barriers to collaboration between closely linked disciplines; e.g., Atmospheric Sciences, Computer Science, Mathematics and Statistics;

TECHNICAL INFORMATION/ISSUES/CHALLENGES

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

- Centralized data repositories and services that link existing and future data systems. For example, a centralized community repository could be created for data submission and sharing.

- Advanced software, tools kits, and services for quality control, in-depth data analysis, visualization, verification, and mining of data (observational and model output). These tools and services need to be user-friendly and accessible by the whole scientific community.

- Common data formats and frameworks for assimilation, modeling, analysis and visualization.

- Common data assimilation framework; currently, each assimilation system uses its own framework for data I/O, processing, and running algorithms.

- Collaboration tools, platforms, and frameworks (e.g., Wiki for data)

- Server-side processing tools for data processing, analysis, visualization

COMMUNITY NEXT STEPS

1. **List of what your community needs to do next to move forward how it can use EarthCube to achieve those goals:**

- A pilot project on coordinated, distributed national ensemble prediction that involves universities that are interested in participating

- Developing a prototype system that links data sets/systems together, involving the most used projects like the reanalysis data sets; Develop a system that works seamlessly, and then expand to include other data sets/systems

- Continued discussion with the goal of developing a concrete plan for greater coordination of ongoing and future programs and facilities that serve the data assimilation and prediction communities, and developing a next-generation testbed facility to advance the science.

- PI meetings to leverage and expand communication, and enhance data sharing, and facilitate sustained interactions

- Entrain current undergraduate and graduate students into research and educational activities related to "big data", ensemble prediction and data assimilation, and EarthCube, to move these initiatives forward for the future scientific workforce.

- Reach out to other geoscience communities, including climate, space weather, oceanography, hydrology, and air-quality, as well as the computer and information science communities.

EXECUTIVE SUMMARY: WORKSHOP RESULTS

Doug Walker, University of Kansas, and Basil Tikoff
University of Wisconsin - Madison: 10/23/12

Earth Cube Workshop Title: EarthCube Domain End-User Workshop for Structural Geology and Tectonics

Introduction: The areas of Structural Geology and Tectonics are at the core of the modern Earth Sciences. For example, the Structural Geology and Tectonics (SG&T) Division is the largest division in the Geological Society of America, and Tectonics is one of the principle foci of the American Geophysical Union. This domain group strives to understand Earth's structural state and deformation processes at all spatial and temporal scales. In addition, much of the motivation for endeavors in the geological sciences is framed in the plate tectonic paradigm.

Workshop participants (35 in total) were all from the US. They represented a variety of disciplines from neotectonics and deformation of the Earth's surface to researchers specializing in ductile deformation. Participants from computer science areas (9%) and students (18%) made important contributions to the effort. The main outcomes of the EarthCube workshop discussions are summarized below.

SCIENCE ISSUES AND CHALLENGES

1. **Important science drivers and challenges:** Participants identified several high-priority science questions that will be the focus of interdisciplinary efforts during the next 5-15 years.

   - What is the evolution of geological structures in three dimensions and at all spatial scales?
   - How can we use the rock record of deformation to better assess the rheology of the crust and upper mantle in different tectonic settings and over different spatial and temporal scales?
   - What are the timescales of different geological processes (fault motion, magmatism, landscape development, etc.) and how do they interact with each other?
   - How do we integrate between short-term (e.g., earthquakes) and long-term (e.g., mountain building) geological processes?
   - How do landscape development and other processes at the Earth's surface relate to geological structures and processes within the Earth's lithosphere?
   - How do mantle processes influence crustal deformation, and what is the dynamic interplay between magmatism, deformation, and mantle flow?

2. **Current challenges to high-impact, interdisciplinary science:** Several themes emerged as consistent challenges faced across disciplines.

   - The Structural Geology community has not adopted conventions for publishing and interchanging digital information, nor has it determined how the data will be archived. Further, primary structural geology data is typically published as derived products. For example, many structural measurements are reported as stereonets, which do not provide information about spatial position, observation quality, or provenance. Consequently, it is difficult for members of the Structural Geology and Tectonics community to collectively share data with each other or even rigidly adhere to NSF's data-gathering requirements. The ability to make primary data (field and laboratory) universally available would be beneficial to both the current and future researchers.

- Researchers and research groups collect data for use as individuals or a small research group, generally using a wide variety of data acquisition workflows and technology. The result of this individualistic approach to data collection is inconsistent formatting, no standard file types, and a complete lack of metadata. A community effort to establish a data standard is needed, together with development of tools to allow multiple data collection schemes to be modified to conform to the data standard for broader data sharing.
- Although researchers are generally willing to share data, data sharing typically requires a trip to the researcher's home institution to gather hard-copy maps and field notes. The ability to access data from diverse sources will greatly enhance efficiency, by allowing researchers to integrate and build on existing knowledge and new results from distant researchers. This approach may require a change in culture within the community concerning an overall willingness to share data.
- Structural geologists typically use a wide variety of data (e.g., field, microstructural, experimental, and modeling) and across a wide variety of disciplines within the geosciences (e.g., geophysics, sedimentology, petrology, etc.) and external to it (e.g., material science, engineering, etc.). Databases in these other fields (e.g., geophysics) typically require both knowledge and computer skills that are often too technical for use by anyone not in that specific subfield. Finally, structural geologists often integrate data over a wide variety of spatial and temporal scales, and numerical modeling is a powerful technique to facilitate this synthesis. Most numerical modeling resources, however, require a significant level of knowledge and background to be used appropriately, which precludes them from being used routinely to test new hypotheses.
- The Structural Geology and Tectonics community has a strong tradition of collaborative science and of community. What is only beginning to be developed, however, is a community of practice to build shared resources and conventions that take advantage of technological advances of the last 20 years. This situation is a result of historical development of structural geology, which has typically relied on researchers working in isolation without needing advanced technology for data collection. Thus, the big challenges are: 1) Community building to support development and adoption of new technology-based approaches to conducting science; and 2) Development and adoption of technology (software and hardware) to enable standardized data interchange by supporting standardized framework for data acquisition and management.

TECHNICAL INFORMATION/ISSUES/CHALLENGES

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

- Workshop participants considered developing conventions and technology for data interchange and documentation to be the highest priority component of cyberinfrastructure needed by the community. This system should be web accessible and allow discovery, access, and reuse Structural Geology data. The scope of such a SG&T Database (or Dataspace) was not developed in detail, although it was recognized that field and microstructural observations would be need to be geospatially referenced. Standards and technology developed by various groups (OpenGeospatial consortium, IUGS Commission for the Management and Application of Geoscience Information (CGI), W3C) were mentioned, and these approaches could be used in the development of such a system.

- There was agreement that analytical tools routinely used to evaluate structural data should be developed in the context of this SG&T Database. These tools include - but are not limited to - stereonet plotting, shape preferred orientation analysis, rotation of data, calculation of finite strain, vorticity analysis, spatial error analysis, three-point problems, etc.. A specific set of new tools would be focused on processing map data. If convenient and powerful tools were available for compiling and analyzing geologic maps and map data, workers would have a natural incentive to use the tools. At the same time, the map tools could serve as a front end for larger map databases. Maps could be designated as "private" until publication, but once public, they would be available to researchers around the world.

- A vast amount of Structural Geology data already exists in the form of geologic maps. These maps contain primary data and are at the very core of the field. Most of these are not in digital form, and the workshop participants considered the digitizing of these legacy data to be very important to the community. This task was considered to be a potentially high impact investment in digital conversion. Semi-automatic to rapidly guided digitizing is considered by the group as an appropriately challenging endeavor for EarthCube. The use of cross-sectional data is particularly challenging, because cross sections involve increased interpretation and their vertical orientation is poorly handled in existing map-based approaches.

- Development of innovative methods to build and visualize interpreted structural histories would be very useful to the structural geology and related geoscience communities.

- Because Structural Geology and Tectonics relies on integration across the Earth Sciences, scientists and students in this area must use data and tools from other fields in the geological sciences. For example, many structural geologists working in neotectonics need access to GPS and LiDAR data. In practice, it can be difficult to find the appropriate data; when found, the user may not be aware of, or how to use, the appropriate tools to solve their structural problems. At a minimum, maintaining a listing of tools and data is critical. More significant advances would involve cataloging resources for best practices and tool use, in addition to making more accessible interfaces for data from other domains.

- There was a keen interest in developing digital laboratory/field notebook software for wide adoption to increase efficiency in the field and facilitate data integration. The concept of the science workbook would be to allow a researcher real-time (or pre-loaded) access to all the geological data from a specific region. This science workbook would form the basic cyberinfrastructure for interacting easily and seamlessly with the database noted above. If well designed and made sufficiently adaptable, the software could be tailored in part to be the front end for the structure database; data collected to be immediately uploaded to the structural geology database (although the data might not become publicly available immediately, to allow for field re-checking, etc.). This software would be platform independent and would have to run on devices from smart phones to pads to tablets to desktops. The development of this type of science workbook would be an important step in developing a cyberinfrastructure for Structural Geology as well as all field-based sciences. Various existing software provides a starting point for defining the functionality and implementation of the science notebook.

EARTHCUBE END-USER PRINCIPAL INVESTIGATOR WORKSHOP

Executive Summary

August 14-15, 2013
Tucson, AZ

Workshop Convener:
M. Lee Allison, Arizona Geological Survey

Workshop Organizing Committee:
Joel Cutcher-Gershenfeld, University of Illinois
Kim Patten, Arizona Geological Survey
Genevieve Pearthree, Arizona Geological Survey
Erin Robinson, Foundation for Earth Science
Steve Richard, Arizona Geological Survey

With Support From:
Rahul Ramachandran
The University of Alabama in Huntsville
National Science Foundation, Award EAR-124039

INTRODUCTION

A two-day 'End User Principal Investigator' workshop was held August 14-15 in Tucson, Arizona, bringing together geoscience domain and cyberinfrastructure scientists who are organizers or participants in the two-dozen EarthCube End-User Workshops. They were joined by a small number of cyberinfrastructure specialists and social scientists, whose purpose was primarily to listen to the presentations and discussions, and provide input on ways to fulfill identified needs in the short and long-term.

The goals of this workshop were to synthesize outcomes from the completed end-user workshops, increase communication among the scientific domains represented, assist with planning for upcoming workshops, and to lay the groundwork for producing documents to inform and guide the EarthCube community, including upcoming End-User Workshop organizers, and NSF EarthCube awards (Test Enterprise Governance, Building Blocks, Conceptual Designs, and Research Coordination Networks).

The 32 in-person and 8 virtual workshop participants represented a wide variety of scientific and technological fields, EarthCube End-User Workshops and EarthCube Working Groups (groups funded to write roadmaps prior to the June 2012 EarthCube charrette):

EarthCube End-User Workshops

1. MYRES (Meeting of Young Researchers in Earth Sciences) V: The Sedimentary Record of Landscape Dynamics
2. Envisioning Success: A Workshop for Next Generation EarthCube Scholars and Scientists
3. Structural Geology and Tectonics
4. EarthScope
5. Calling All Experimentalists- Experimental Stratigraphy
6. Shaping the Development of EarthCube to Enable Advances in Data Assimilation and Ensemble Prediction
7. Engaging the Critical Zone Community to Bridge Long Tail Science with Big Data
8. Envisioning a Digital Crust for Simulating Continental Scale Subsurface Fluid Flow in Earth System Models (Hydrology/Dark Data)
9. Modeling
10. Cyberinfrastructure for Paleogeoscience
11. Education
12. Community-based Cyberinfrastructure for Petrology, Geochemistry, and Volcanology
13. Sedimentary Geology
14. Integrating the Inland-Waters Geochemistry, Biogeochemistry and Fluvial Sedimentology Communities
15. Deep Seafloor Processes & Dynamics
16. Integrating Real-Time Data Into the EarthCube Framework
17. Articulating Cyberinfrastructure Needs of the Ocean Ecosystem Dynamics Community
18. Increasing Access to and Relevance of Marine Seismic Data
19. Bringing Geochronology into the EarthCube Framework
20. Rock Deformation and Mineral Physics Research
21. Community-Based Cyberinfrastructure for Polar Science Instrumentation, Technology, and Environmental Monitors

EarthCube Working Groups

1. Cross-Domain Interoperability
2. Data Discovery, Mining and Access
3. Earth System Model
4. EarthCube Stakeholder Alignment Survey
5. Governance Framework
6. Physical Samples as Part of Cyberinfrastructure
7. Workflow

A representative from each of the previous workshops summarized the science drivers, challenges, needs, and action items identified in previous workshops. We have synthesized these results by categorizing the various factors identified, empirically defining facets and categories in those facets, and then grouping in the categories. The following section is a discussion of this synthesis.

PREVIOUS WORKSHOPS: SUMMARY OF RESULTS

Science Drivers

Science drivers identified by the user workshops have been categorized to synthesize the results. Dynamic Earth system models are central to science drivers identified by 11 workshops, science methodology related drivers were identified by 6 workshops, and development of shared 3-D Earth models was identified by 5 workshops. Other science drivers mentioned are related to education, biology, new discoveries, planetary models and space weather. Science drivers for each category are presented in detail below.

EARTH SYSTEM MODELS

Many of the issues in critical zone science, climate modeling, hazard assessment, and anthropogenic impacts require more accurate dynamic Earth system models that have greater spatial and temporal resolution, and account for more of the non-linear feedback processes in the Earth system. These models require improvements in modeling algorithms, better understanding of the dynamics of coupled solid-earth, hydrosphere, atmosphere systems, increased computational capabilities, and more accurate and complete spatial and temporal observation data for the state variables that characterize the system.

SCIENCE METHOD

Drivers related to scientific methodology were focused on computational approaches to improved modeling capabilities, optimizing sampling and observation strategy, knowledge representation, and linking scientific results to decision-making processes. The most challenging modeling issues appear to be related to integrating coupled models at different temporal and spatial scales for systems with varying degrees of coupling, and integrating results from ensembles of models (particularly for weather). The limits of model accuracy, both from a computational point of view, and in light of the natural variability of the Earth system need to be better understood. The high cost of obtaining observational data from samples and sensors to understand processes and to enhance dynamic Earth models is motivating investigations of optimized sampling techniques, particularly for field experiments that rely on real-time data. Analysis of events captured by real-time sensor networks requires data and derived products to be presented in a timely fashion and in a form that is useful to decision-makers. More sophisticated approaches to knowledge representation in computational systems is needed to facilitate integration of multi-scale, multi-domain data, and to make results more readily accessible for real-time analysis and understanding, both as part of the experimental process and to improve utilization of scientific information in decision making.

SOLID EARTH MODELS

The need for high-resolution three-dimensional models of the present state of the solid Earth are implicit in science drivers related to tectonics and a better understanding of Earth history. We cannot experimentally test tectonic hypotheses for Earth dynamics because the time scales, temperatures, pressures, and chemical environments of the Earth's interior cannot (for the most part) be reproduced in the lab. Much of our understanding of the Earth is based on abductive reasoning supported by geologic mapping on the surface, borehole data in the near surface crust, and by probing the Earth's interior with various geophysical techniques. An Earth Model resource that supports registration and integration of 3-D models from the community in a dynamic information system enabling feedback, updates, multiple interpretations, and provenance tracking is envisioned as a foundation resource for all kinds of geoscience research. Such a resource would improve reproducibility, reduce duplication of effort, and streamline project development through more straightforward access to available legacy research.

MISCELLANEOUS

Not surprisingly, there were other science drivers identified by various workshop participants that reflect the widely varying interests of the community. Education was brought up as an issue at several workshops, highlighting the importance of cultivating the next generation of cyber-savvy Earth Science researchers. The issues identified centered on how to teach skill sets and habits of mind necessary to use data and models, address novel and ill-structured problems, work in collaborative teams, adopt appropriate data and analysis strategies, and effectively communicate. Some other miscellaneous drivers included understanding the physical limitations for evolution of life, understanding planetary formation, prediction of large solar flare events, and discovery of new, unexpected phenomenon.

Challenges

The top challenges identified by previous workshops related to data (15 workshops), standards (11 workshops), and incentives (7 workshops). Challenges related to data can be categorized into curation, integration, access and documentation.

Data curation challenges (9 workshops) are associated with the process of preparing data for archive and reuse, including transformation into archive formats, documentation, placing it into a secure yet accessible repository, and curation of physical resources such as samples. Many of the 'dark data' issues stem from the difficulty that individual researchers, operating on very limited budgets, experience in trying to curate data produced by their research. Specific challenges include lack of tools, conversion to digital formats, and lack of clearly defined best practices.

Data integration (7 workshops) challenges address processes and capabilities to transform multiple datasets into formats and schema so they can be used together. Issues include data heterogeneity, lack of documentation (provenance, quality, semantics...), different spatial reference systems, and integrating data collected at various spatial and temporal scales or with different sampling strategies.

Data access (9 workshops) challenges relate to processes and capabilities required to allow data consumers to get and use data, including repository architecture, data services, real-time access, definition of information exchanges, and registration of resources in catalogs. Some specific challenges mentioned include tool complexity, access restrictions, bandwidth limitations and difficulty exporting data from databases.

Data documentation (7 workshops) challenges are related to specifications and practices for documenting the content, format (schema, encoding), semantics, provenance, quality, and stewardship of resources. Such documentation is essential to enable cross domain use of data, or reuse/repurposing of data obtained from repositories. Specific issues mentioned include reproducibility of model results, ability to assess uncertainty, difficulty discovering and evaluating data for research purpose, and citing the origin of data.

Standards issues relate to development and adoption of community conventions for data management, documentation, exchange, and analysis. To utilize non-standard, heterogeneous data from different sources requires significant effort to analyze each dataset to understand its content and determine how to transform it to integrate with other data. Lack of standard vocabularies for specifying data schema and property values complicates the problem because the meaning of the data may be unclear. Inconsistent practices for data sharing make each new data acquisition a time consuming learning experience.

Incentives challenges relate to social and financial factors that motivate good data management practices and the sharing of data or models. Lack of credit for data publication, either in the form of citation or career advancement is a major factor. Cost, effort, technical barriers, and concern about misuse outweigh the tangible benefits, especially for tenure track and project-funded researchers.

Some other challenges mentioned include misuse of data, communication between researchers in different domains, cost, absence of data, connecting target users with relevant data, lack of training, and insufficient computational capabilities.

Needs

The most widely identified needs are related to data (12 workshops). As with the challenges, the needs are grouped in to access, discovery, curation, and integration categories. A common thread was the need to make these operations easier and less time consuming, and the inclusion of a broad spectrum of resources included not only data, but models, workflows, samples, and tools. Data access needs (10 workshops) also mentioned the ability to subset large datasets, cross-domain data access, and integrating data access seamlessly into processing workflows. Discovery needs (6 workshops) identified included a single catalog for searching across the gamut of geoscience domains, map-based, space-time ('4-D') searching, and 'smart' searching technology to improve search efficiency. Related to data curation (8 workshops), the need for incentives or change in culture to promote better data management and data sharing was highlighted. Data integration needs (6 workshops) mentioned the need to integrate data in space or time, and to develop better approaches to integrating noisy or sparse data.

A common thread through many of the identified needs was the need for better software tools (10 workshops) to facilitate data documentation, archiving, and publication, data discovery and evaluation, data integration, and data analysis. Data visualization tools in particular were mentioned in issues raised by 7 workshops. In the data integration arena, improved capabilities for subsetting large datasets, transforming between schema, up-scaling and down-scaling, and resampling were mentioned.

The needs for standards (7 workshops) included mentions of standards for web interfaces to data, metadata and data formats, and standardized software libraries. Education needs were prominent (6 workshops), both to develop the workforce to support cyberinfrastructure, and to inform scientists about available technology and how to use it. Also mentioned were the needs to link data with other resources (datasets, samples, projects, researchers, and tools), shared domain databases, improved communication and social networking, governance to coordinate programs and pilot projects, develop metrics, and identify gaps, and the need for funding.

Action Items

The suggested action items from the workshops focused on community building (10 workshops) and communication (6 workshops). Community building activities mentioned include increasing participation in EarthCube, providing online social networks and community collaboration tools, outreach events, codathons, workshops, developing data systems for communities that currently lack such systems, and promoting the emergence of small, targeted workgroups. Communication enhancing activities included a possible journal of geoscience data, tracking of NSF solicitations, listings of scientists interested in sharing data, and facilitation of social networking.

Education-related action items (6 workshops) include developing and testing learning activities for utilizing data and models and promoting research on how humans learn using data and models, and organizing training workshops and webinars on data management, documentation, and other data management and utilization topics.

Data related action items (5 workshops) include implementing linkages from publications to data and models, forming workgroups to develop standards and recommend best practices, compiling inventories of existing resources, and constructing infrastructure to support access to sensor networks and real-time data streams. Specific software processing capability action items were mentioned in 4 workshops, and include data analysis software seamlessly linked to data centers, tools for community annotation of resources, and leveraging existing library and visualization software. Other action items mentioned include implementing a governance mechanism to prioritize resource allocation, and developing a community database for 3-D Earth structure.

PREVIOUS WORKSHOPS: BEST PRACTICES AND LESSONS LEARNED

Representatives from five upcoming EarthCube End-User Workshops (Geochronology, Ocean Ecosystem Dynamics, Marine Seismic Data, Rock Deformation, and Polar Science Instrumentation) gave a short presentation on the goals and target communities of their workshops, and participated in a 'reverse panel,' in which they asked organizers of previous end-user workshops about best practices and lessons learned. Here is a summary of the observations from previous workshops.

1. **Invited Speakers:** Get motivated speakers with a clear, strong vision to open up the discussions, and have them stay throughout the workshop so that they can contribute to workshop dynamics.
2. **Communication:** Ask other Workshop PIs for copies of their proposals, and/or agenda materials to use as examples. Attend another end-user workshop, if possible.
3. **Engaging your respective community:** Workshop steering committee members should reach out personally to potential workshop participants. Personal contact is much more effective than e-mail to get people excited about the workshop and to commit to participating. Potential

personal outreach avenues include meetings, department hallways, and a personal phone call. If someone is unable to attend, ask if they can recommend a colleague.

4. **Workshop Registration:** Develop a Google Form for workshop registration that you can monitor to get a sense of the demographics of people who have registered. For example, if you ask participants for their primary and secondary scientific interest, you can monitor how well your current participant list is covering the necessary disciplines and target invitations to representatives of communities that are not well represented. An example signup form can be found here.
5. **Workshop Scheduling:** In order to maximize in-person participation, do not schedule your workshop for a Monday or for the day after a holiday.
6. **Inviting NSF Program Officers:** Invite multiple NSF Program Officers (POs) to attend your workshop**.** The POs will be there as observers and will send a signal to the domain scientists that many programs within the NSF are supporting EarthCube.
7. **Workshop Action Items:** Identify use cases and specific, actionable next steps, such as a RCN (Research Coordination Network) proposal, workshop, white paper, or other concrete action items.
8. **Discussion & Breakout sessions:**
    a. *Minimize the talks/presentations and devote more time for discussion.* A good example is the CZO (Critical Zone Observatories) workshop. With a big chunk of time – 1.5 hours – the CZO workshop organizers and participants came together on a proposal that was submitted only four weeks after the workshop and was just recommended for funding.
    b. *Highlight the breakout discussions.* For many workshops, the breakout discussions in small groups were the most productive aspect of the workshop. Make sure a scribe and a facilitator are identified for each break out to capture discussion and conclusions.
    c. *Debriefing:* Make sure you schedule enough time for the breakout sessions to debrief with the whole group, in order synthesize and integrate outcomes of each of the breakout sessions.
    d. *It is important to get people thinking big.* Allow people to air their immediate frustrations and workflows, but also encourage them to think bigger and look beyond their immediate issues. A good starting question is *what would you do if CI was not a barrier?*
9. **Virtual Participants:** A robust virtual participation component should be enabled. Components include good microphones, sound system and visual display, so that virtual participants can see the presentations and participate in group discussions and breakouts.
    a. *Sound Quality:* Place microphones around the room that connect to the virtual participants. If necessary, have microphone stewards who move mobile microphones around so that the speaker is always audible for virtual participants. In one of the workshops Polycom systems were purchased on Ebay and there was a projector in every room. Virtual participants won't be able to participate if they can't hear the discussion.
    b. *Breakout Sessions*: Integrate virtual participants into in-person breakout groups instead of assigning them to their own virtual breakout session.
    c. *Webex Capabilities*: EarthCube has a Webex account with three host licenses that can be used to provide a virtual participation component for the End-User Workshops. Up to the three meetings can occur simultaneously on the EarthCube Webex (one per host license). Each host license can accommodate up to 100 people. If you'd like to use the Webex account, please contact genevieve.pearthree@azgs.az.gov.

10. **Note-Taking**
    a. *Use Two Screens*: Have two screens and two projectors set-up. One screen will display the presentation material; the other will display notes taken in real-time.
    b. *Flip charts:* For breakout sessions, have participants take notes on a Google doc (for the record) and on a flip chart (for people in the room). Make sure you have one flip chart per breakout session.
    c. *Use Google Docs to take notes.* Google docs worked very well for many of the workshops. Take a look at some examples from other workshops: Critical Zone, Real-Time Data, Paleogeoscience, Inland Waters Geochemistry.
    d. *Training:* All workshop participants should have access to and know how to use Google docs prior to the workshop. You might need to train them prior to the workshop (i.e. a webinar or something similar).
    e. *Accessibility:* Make sure all the Google docs that will be used for note-taking are created prior to the workshop and that they are easy to find. On the agenda, share the link to the Google doc that corresponds to each activity so that it is easy to find. Make sure the Google docs are editable by all workshop participants so that many people can contribute and correct notes in real-time.
    f. *Note-taking Protocol:* Establish a note-taking protocol and identify a principal note-taker for each breakout session and group discussion. For example, in the CZO and Real-time data workshops, the note-taker's cursor was on top. Other people could add entries down below that the note-taker could then integrate into the main notes.
    g. *Designate two note-takers for each breakout session*: One person will take notes on a flip chart, the other can record the notes on Google docs.
    h. *Other Technology.* Tablet computers also work well for taking notes during breakout sessions.

11. **Special Events:** Arrange a tour, use live displays, or arrange another special event during the workshop to spark discussion and encourage attendance.

12. **Workshop synthesis & final report**: The leaders and scribes of the break out groups and the organizing team should get together for a half day after the workshop to synthesize workshop results before workshop organizers leave. This will prevent a long delay in finishing the workshop executive summary. In advance of the workshop, conveners should get commitments from other workshop organizers to stay late to synthesize workshop results.

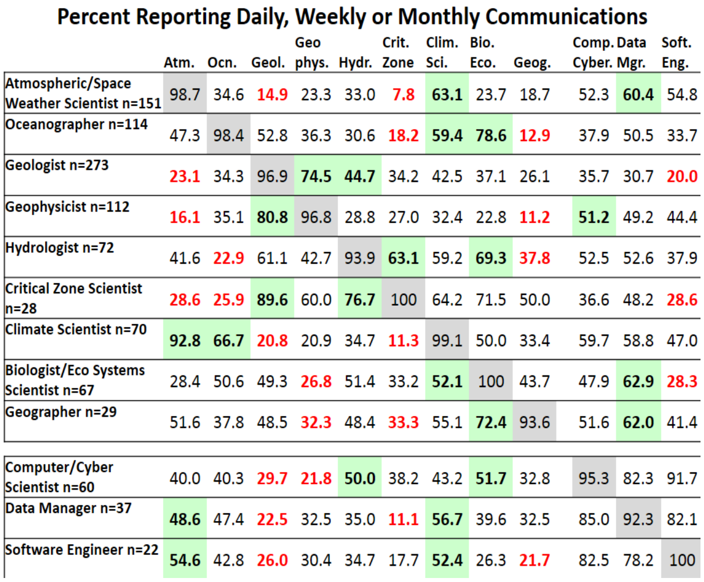
13. **Workshop Survey:** Encourage workshop participants to take a survey on the workshop, noting what went well, what could be improved, and what next steps they'd like to take. For an example, please see the EarthCube End-User Principal Investigator Workshop Follow-Up Survey.

The Future of EarthCube

Potential Paths Forward

Workshop participants identified and discussed five potential paths forward for EarthCube. For each of these strategies, breakout groups analyzed the current state, desired/future state and delta state (action items and milestones needed to reach the desired state).

1. **Low Hanging Fruit:** *How do we leverage and expand upon existing communication and collaboration?* Focus on items for which there is already some convergence on requirements, ideally with existing software systems in place. The target communities are indicated by high percentages of researchers reporting interactions between them (Figure 1).



Percent Reporting Daily, Weekly or Monthly Communications

	Atm.	Ocn.	Geol.	Geo phys.	Hydr.	Crit. Zone	Clim. Sci.	Bio. Eco.	Geog.	Comp. Cyber.	Data Mgr.	Soft. Eng.
Atmospheric/Space Weather Scientist n=151	98.7	34.6	14.9	23.3	33.0	7.8	**63.1**	23.7	18.7	52.3	**60.4**	54.8
Oceanographer n=114	47.3	98.4	52.8	36.3	30.6	18.2	**59.4**	**78.6**	12.9	37.9	50.5	33.7
Geologist n=273	23.1	34.3	96.9	**74.5**	**44.7**	34.2	42.5	37.1	26.1	35.7	30.7	20.0
Geophysicist n=112	16.1	35.1	**80.8**	96.8	28.8	27.0	32.4	22.8	11.2	**51.2**	49.2	44.4
Hydrologist n=72	41.6	22.9	61.1	42.7	93.9	**63.1**	59.2	**69.3**	37.8	52.5	52.6	37.9
Critical Zone Scientist n=28	28.6	25.9	**89.6**	60.0	**76.7**	100	64.2	71.5	50.0	36.6	48.2	28.6
Climate Scientist n=70	**92.8**	**66.7**	20.8	20.9	34.7	11.3	99.1	50.0	33.4	59.7	58.8	47.0
Biologist/Eco Systems Scientist n=67	28.4	50.6	49.3	26.8	51.4	33.2	**52.1**	100	43.7	47.9	**62.9**	28.3
Geographer n=29	51.6	37.8	48.5	32.3	48.4	33.3	55.1	**72.4**	93.6	51.6	**62.0**	41.4
Computer/Cyber Scientist n=60	40.0	40.3	29.7	21.8	**50.0**	38.2	43.2	**51.7**	32.8	95.3	82.3	91.7
Data Manager n=37	**48.6**	47.4	22.5	32.5	35.0	11.1	**56.7**	39.6	32.5	85.0	92.3	82.1
Software Engineer n=22	**54.6**	42.8	26.0	30.4	34.7	17.7	**52.4**	26.3	21.7	82.5	78.2	100

2. **Undervalued Opportunities:** *How do we close the big communication/collaboration gaps?* Focus on communities that do not currently report much interaction. Rapid progress may be possible by engaging previously separate groups (Figure 1).

3. **Institutional Stakeholders:** *How do the big players (data centers, academic centers, government agencies, etc.) work together?* Work with existing data centers that already have signification data holdings and experience, integrate those practices, and build from there.

4. **Making Dark Data Light:** *Can the transformational goals of EarthCube be realized if we don't bring dark data to light?* Capture legacy (dark) data (in publications, not accessible in databases) and grey data (data in researchers' files, never edited and documented for sharing). Emphasis on 'long tail' ('mainstream') community.

5. **Next Generation of Geoscientists:** *Emphasize education and workforce training*. Train the next generation in the use of cyberinfrastructure; they will be more active users and are better suited to determine priorities.

Vision of Success

From these five potential paths forward, workshop participants elucidated a vision of EarthCube success. In this vision, EarthCube is a cyberinfrastructure for geoscience research that fosters new, transformational science using formerly unrelated resources by enabling simple discovery, evaluation and access to all data. EarthCube will be an easy-to-use system, similar to the World Wide Web, and provide useful capabilities to a broad spectrum of geoscience users. EarthCube will align scientific needs and technology development, building on existing cyberinfrastructure and geoinformatics while embracing open source culture and methodologies. The architecture required to meet science requirements is envisioned as emerging from convergence on conventions for services, interchange formats, and protocols. Geosciences communities without a mature cyberinfrastructure framework will develop a more mature framework. A successful cyberinfrastructure will enable broad participation in science, engaging the academic community as well as public citizen science.

Specific short and long-term components of an EarthCube vision of success are outlined below, organized according to their association with governance, science practice, software development, community building, and education.

1. **Governance: Co mmunity adoption of policy and specifications**
    a. Short-Term. Achieve consensus on requirements to increase the efficiency (cost/benefit ratio) and effectiveness of workflows for finding, sharing, and reusing data, tools and models. Use these requirements to establish infrastructure development priorities. Governance processes foster best practices for collecting, monitoring, and utilizing community feedback on the quality of content, reuse of resources, including practices and policies for access and authentication for this information, as well as privacy concerns

    b. Long-Term. A portfolio of community specifications is in place, under stewardship of EarthCube governance, and in use by the community.

2. **Science Practice: Changing the culture of how scientists work**
    a. Short-Term. Researchers frequently find, share, and use data, tools, and models from the EarthCube system in ways that would not have happened otherwise. Data, tools, and models are being 'mashed up' in new combinations to support new scientific discovery.

    b. Long-Term (Data). EarthCube supports everyday data discovery, access, reuse and stewardship with tools integrated into scientist's normal workflow. Scientists use and reuse data that is easily accessible through EarthCube repositories. Intelligent tools allow people to find, understand, and use data from disciplines outside their area of expertise. A culture of data sharing and availability based on EarthCube technology promotes sharing of published modeling tools, data, results, and big data.

    c. Long-Term (Scientific Discovery). EarthCube facilitates scientific discovery and cyberinfrastructure evolution for interdisciplinary science, particularly among Early Career scientists.

    d. Long-Term (Standards Process). EarthCube fosters standardized data collection procedures and descriptions across communities with requirements for formats and conventions. It is easy to create and publish useful metadata to enable data reuse.

3. **Software Development: New software capabilities and their adoption for daily use**
    a. Short-Term. Prototype systems are operational, based on currently functional cyberinfrastructure components, and provide access to essential variables and operations via web APIs (Application Programming Interfaces). These prototypes demonstrate integrated, interoperable geospatially searchable data with mapping capabilities for visualization, integrative tools for synthesis and analysis, such as a “digital crust” or “critical zone” prototype demonstration and tools for curating 'dark' (legacy) data, 'grey' data, and continuously created data.

    b. Long-Term (Entry point). An online, 3-D virtual globe provides an entry point to explore Earth Science data at different resolutions and from different perspectives.

c. Long-Term (Platform). This EarthCube platform provides community cloud cyberinfrastructure with APIs for data storage, resource cataloging and discovery, and high performance computing that Earth scientists use as a foundation for application-specific user interfaces and tools both within and across communities. Platform components support curating resources in ways that are immediately useful for others to reuse (understand and access), data management practices that maximize resource documentation and minimize the overhead for Earth scientists collecting data or generating new resources, and support integrating data with spatial and temporal scaling of essential variables for input to large, complex models or real-time use.

4. **Community: Development of an EarthCube community of practice**

   a. Short-Term. Alliances among institutional data centers implement pilot interoperability and integration experiments. Communication between geo scientists and cyber scientists exploits and promotes synergies between individual domains. Metrics are in place that demonstrate progress towards realizing the EarthCube vision.

   a. Long-Term. The Earth Science community identifies itself through participation in EarthCube, as data providers, data consumers, system developers, maintainers, and managers. EarthCube participants move from a set of individuals working on a global platform to a set of individuals who are the global platform.

5. **Education: Training and education to foster cyberinfrastructure use and adoption**

   a. Short-term. EarthCube provides intuitive, modular learning objects and self-directed lessons that are used by teachers from K-12 through the graduate level.

   b. Long-Term. An online, EarthCube 3-D virtual globe is the entry point for students to explore Earth Science data at different resolutions and from different perspectives. EarthCube helps everyone (including individuals outside of the geosciences) better understand how to use and interpret data, because EarthCube helps improve computational literacy at all levels.

EarthCube Fears

The following fears for the future of EarthCube emerged during discussion at the workshop.

1. **Adaptability**. Workshop participants are concerned that the data management backlog grows faster than our ability address cultural obstacles to data sharing.
2. **Funding**. Workshop participants are concerned that NSF funding won't materialize to help scientific domains mature their cyberinfrastructure framework, that NSF funding and support will not continue into the future, and that there is a lack of coordination between EarthCube Program Officers and other Program Officers within NSF.
3. **Governance and Prioritization**. Workshop participants are concerned that the EarthCube community will be unable to agree on priorities and demonstrate concrete progress. The community will have to establish relative importance of building innovative cyberinfrastructure (software, tools, etc.), bringing 'dark data' to light, and standards development, and use these priorities to guide resource allocation.
4. **Community Engagement**. Workshop participants are concerned that EarthCube will not engage a broad community across the geosciences.

5. **Data practices.** Workshop participants are concerned that data quality control will not be built into the system. Thus, processes must be developed and implemented to minimize data misinterpretation or misuse; access controls and respect for data ownership are necessary to assure that data are not shared too soon or too widely; adequate credit must be given for contributing data and models.
6. **Utility of EarthCube.** Workshop participants are concerned that EarthCube will not produce something useful to wide segments of the community. If we attempt to build cyberinfrastructure that meets the needs of all, we might end up with cyberinfrastructure that is ineffective at meeting anyone's specific needs. There must not be so many data and tool choices that data consumers are overwhelmed. EarthCube should enable global science, not solely focused on the U.S.A.
7. **Education and Workforce Training.** Workshop participants are concerned that education and training will be left out. Thus, EarthCube should foster training in data, modeling, and information science and technology to meet the needs of geoscientists.

MOVING FORWARD: ACTION ITEMS

A variety of near-term action items were established for workshop participants.

1. **EarthCube White Paper:** Participants will collaborate on an EarthCube white paper to serve as the voice of the end users at this key juncture for EarthCube (to be submitted to *Eos* or another appropriate forum, target date is October 2013).
2. **Workshop Meeting Report:** Participants will collaborate on a 500-word *Eos* Meeting Report, to be submitted to *Eos* as soon as possible.
3. **Data Facilities End-User Workshop:** An End-User Workshop for Data Facilities will be organized in January 2014 in conjunction with the Facilities Assembly Group as part of the upcoming EarthCube Test Enterprise Governance process.
4. **Low-Hanging Fruit:** Workshop organizers were tasked with identifying one or two 'low hanging fruit' projects for their scientific community, which will be compiled and presented as recommendations to NSF.
5. **NSF and Governance Webex Presentations:** A series of Webex presentations to the EarthCube community will be held in the fall of 2013. This first of these presentations will be an NSF announcement of funded EarthCube components, followed by an introduction of the EarthCube Test Governance Process and announcement of opportunities for continued participation in EarthCube, such as the EarthCube Test Enterprise Governance Assembly and other avenues.
6. **EarthCube Events at Professional Conferences:** Participants will collaborate with NSF to hold an EarthCube event, such as a Town Hall, at the Geological Society of America (GSA) 2013 Annual Meeting and the American Geophysical Union (AGU) 2013 Annual Meeting, in addition to organizing EarthCube events at other professional conferences in the geosciences.

Increasing the Access to and the Relevance of Marine Seismic Data

December 11-13, 2014 – San Francisco, CA

Convener: James A. Austin, Jr., University of Texas at Austin, Institute for Geophysics (UTIG)
Steering Committee (alphabetical order): David Arctur, University of Texas at Austin (designated EarthCube liaison); Nathan Bangs, UTIG; Suzanne Carbotte, Lamont-Doherty Earth Observatory (L-DEO); Jon Childs, U.S. Geological Survey (USGS); Adrian McGrail (ION); John Snedden (UTIG)

Executive Summary

Introduction

This NSF-supported workshop brought together knowledgeable U.S. and international representatives of the marine geology and geophysics academic community, along with selected members of the offshore hydrocarbon business, representing both key industry data users and seismic data vendors. During the workshop, these participants, both in plenary sessions and in successive breakout configurations of smaller groups, considered the following primary objectives and related relevant topics:

- Review of some successful examples of academic use of preexisting industry marine seismic data.
- Discussion of strategies for developing joint industry-government-academic acquisition of new data, which are consistent with commercial/academic imperatives and data restrictions.
- Discussion of the potential for a revised model for research vessel *Marcus G. Langseth* operations.
- Link to (new NSF initiative) EarthCube: Identifying current limitations to the conduct of marine seismic research in terms of data management, processing, analysis and visualization. Envisioning the future of the science and identifying scientific challenges and data/cyberinfrastructure for the next decade.
- Exploration of more efficient use of U.S. and non-U.S. national seismic databases. How can international academic communities work together to optimize seismic data resources and related cyberinfrastructure? Identification of seismic data resources unfamiliar to the academic community.
- Codification of industry sources of 2D/3D seismic data for future U.S. academic community use, *e.g.*, in support of scientific ocean drilling efforts in areas where both academia and industry will benefit from such drilling.
- Construction of a roadmap to augment existing models of seismic data access for the U.S. and international academic communities.

Of the 51 participants, 39 were from academia; we targeted various levels of seniority. In addition, 12 were either from industry (petroleum regulation, petroleum exploration, engineering, software) or from U.S. government agencies (USGS, Bureau of Ocean Energy Management [BOEM], National Oceanic and Atmospheric Administration [NOAA]). Four program managers attended from the National Science Foundation (NSF); active participation by these managers in the ongoing discussions contributed to the workshop's success. Finally, 4 participants came from outside the U.S. (Norway, Germany [2], Japan). (For a complete list of participants and their affiliations, see Appendix 1.)

The workshop took place over ~2.5 days just prior to the 2014 Fall Meeting of the American Geophysical Union. The originally envisioned Agenda is included as Appendix 2. Presentations given at the workshop are the property of the presenter; some data shown were proprietary. These may be requested by contacting the presenter (see e-mail addresses with participants list, Appendix 1). However, a city-wide power failure associated with an intense storm during virtually all of Day 1 forced a complete re-thinking of the original agenda. As a result, break-out discussions in several groups took place during that day, and most presentations in plenary sessions scheduled for Day 1 took place instead during days 2 and 3. Fortunately, all participants rose to the challenge, and very effective communication occurred (see summaries below).

Break-out group discussion topics and summary action items:

o **Strategies for developing joint industry-academic acquisition of new data, based upon known successful case studies, which are consistent with commercial and academic imperatives and data restrictions**

Discussion leaders: Childs (USGS) and McGrail (ION)

o **Construction of roadmap to augment existing models of seismic data access for the U.S. and international communities**

Discussion leaders: Childs (USGS) and Damm (BGR)

(For a summary of the notes from these two break-out sessions, see Appendix 3.)

Action items:

1. Develop a roadmap outlining a Joint Industry Project (JIP) to analyze shallow geomorphology using "non-sensitive" (i.e., upper 1 second of seismic data record) of industry 2D/3D seismic in the Mississippi Canyon, Gulf of Mexico (GoM): Torry, Prather, Reece, Sager. Timeline: Q1, 2015. (For a more complete summary of this item prepared by Prather, see Appendix 4.) First meeting held in February, 2015, in Houston; follow-up meeting proposed for 2Q 2015. Prather to lead.

2. Schedule EarthCube presentations at selected professional society meetings (AAPG, SEPM, AGU, OTC): Ransom. Timeline: calendar 2015.

3. Develop increased interoperability between both public-domain and industry databases (*e.g.*, NGDC, AGI, NAMSS) through Web services: Childs, Carbotte, Jencks. Timeline: calendar 2015.

o *Discussion of a revised model for research vessel Langseth operations*

Discussion leaders: Sawyer (Rice) and Bangs (UTIG)

(For a summary of the notes from this breakout session, which spanned days 1 and 2, see Appendix 5.)

Action Items:

1. MLSOC and the operator at L-DEO could develop a web-based roadmap for countries that need support for planning and executing a cruise on *Langseth*; the ship and scientists needed for analysis and interpretation (from the U.S.?) could be pitched as a "package". Timeline: calendar 2015.

2. Solicit periodic “Letters of Intent” from the U.S. academic community. A committee (membership TBD, by MLSOC) would vet these letters; set up multi-year *Langseth* future general directions that follow a regional model of cruise planning. Encourage a separate NSF panel for judging *Langseth* proposals relative to each other (with MLSOC?) Timeline: calendar 2015.

3. MLSOC prepares to be able to provide science consulting (*e.g.*, a workshop for new users, as has already been initiated by MLSOC annually at the AGU Fall Meeting); some members of the community seek *Langseth* knowledge from colleagues, but others (*i.e.*, new users) do not know who to talk to about preparing proposals to use the vessel. Timeline: calendar 2015.

4. MLSOC should update the 2010 Incline Village workshop “flashy document.” (*Marine Seismic Imaging: Illuminating Earth’s Structure, Climate, Ocean and Hazards)*. Timeline: calendar 2015.

5. MLSOC could poll the relevant community to find out which *Langseth* “tools” are most important. Timeline: calendar 2015.

6. Improve advertising of *Langseth* and the data she collects, and improve the educational footprint; MLSOC with the L-DEO science operator could develop a website for these purposes. Timeline: calendar 2015.

7. Endorse “training cruises”; reserve bunks for early career scientists (*being done*).

o **Discussion of strategies for developing joint industry-academic acquisition of new data, which are consistent with commercial and academic imperatives and data restrictions.**

Discussion leaders: Snedden (UTIG) and McGrail (ION)

(For a summary of the notes from this breakout session, ref. also Appendix 3, see Appendices 7 and 8.)

Action Items:

1. Expand/make more dense (*i.e.*, reduce receiver spacing) long-offset OBS seismic refraction collection; eventual formation of a JIP?

2. Independent testing of seismic source arrays to evaluate marine wildlife impact.

3. Consider bringing students onboard industry seismic vessels, for training purposes.

4. Formation of a committee to develop a road map or actual JIP to make use of shallow sea bed seismic mages to understand deep-water processes (see more detailed action item description on this above).

o **Links to EarthCube: Science Drivers and Challenges** *(Note: This breakout group focused on science drivers and challenges, with less emphasis on actionable items. The relationship of seismic imaging within the NSF EarthCube initiative will be explored with a follow-up effort to fund a Research Coordination Network (RCN) to advance the interests of the seismic imaging end-user community within the currently evolving EarthCube framework).*

Discussion leader: Arctur (UT/Austin)

While the petroleum exploration industry is primarily interested in marine seismic surveys to understand where oil and gas may be found along the Earth's continental margins, the academic community wishes to study broader issues of stratigraphic and structural evolution, including rifting and thermal subsidence, long-term sea-level history, magmatism and geochemistry, geomorphology and earthquakes. The industry participants at the workshop expressed strong philosophical support of academic science research, but issues of substantive collaboration and funding remain.

Important science drivers and challenges**:** Participants identified several high-priority science questions that will focus interdisciplinary efforts during the next 5-15 years:

A. *Digital Earth: get enough data with sufficient resolution and history of change, from which to understand/forecast/predict:*

- Major subsidence events/earthquakes/volcanic eruptions/hazards;
- Dynamics of sea level rise; likely shoreline changes over next 100 years;
- Combinations of factors that control the modes of slip on fault zones;
- Combinations of factors that control the modes of rifting in continental crust, especially to break-up and seafloor spreading; this includes the architecture of magmatic systems and dynamics of magma transport and associated eruptions;
- Sedimentary evolution of continental margins from Greenhouse to Icehouse (warm climates to cold climates);
- Visualization of Earth systems, from the scale of tectonic plate geodynamics to reservoir simulations.

B. *Improved monitoring systems, long-term and real-time, with advanced simulation:*

- Monitor crustal displacement and microseismicity; monitoring with active source time-lapse (4D) studies;
- Anticipate major seismic events;
- Develop and maintain a fleet of small autonomous seismic sources and seismic detectors that could be monitored with intelligent drones to acquire seismic field data safely without harming the environment.

C. *Understanding impacts to marine life, esp. endangered species (i.e., settle the ongoing environmental debate):*

- Marine mammals, etc.;
- Are chemosynthetic seafloor communities (e.g., those adjacent to fluid seeps in basins like the GoM) affected by seismic surveys?

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

A. *Should we seek development of a national academic data computation capability for processing field data?*

- Strengthening academic HPC could do this, but it's controversial; industry tends to do this better due to extensive familiarity with ground-truth; *e.g.*, noise correction is highly variable & requires expert geologic knowledge of target areas. It's not just about "having HPC."
- Alternatively, we (EarthCube) could work to develop licensing schemes for commercial tools, data and use in academic centers, but this is institutionally difficult for industry.
- One area of development would be to strengthen industry/academic relationships (outreach: share lessons learned, contractual experiences, institutional best practices).

- Some from industry suggested academia could bring valuable software engineering skills to bear on processing techniques to use HPC wisely, and work with industry to incorporate domain expert knowledge.
- Academia could help focus on visualization tools to look for patterns in imagery, relationships with other variables, map layers, etc.

B. Discovering sources of relevant, accessible data:

- We are building the ability to do federated search of catalogs of marine seismic data resources, via GeoMapApp and GEOSS. Also need access to academic, government agency and international data resources.
- GEOSS (geoportal.org) provides a single point of search to dozens of community catalogs for Earth science, such as NASA GCMD, Pangaea.de
- Could register additional major catalogs in GEOSS
- Look into Microsoft LayerScape

C. Data preservation, long-term archiving and access:

- UTIG/L-DEO/NGDC have systems and practices in place - collecting all possible (data) sources for long term redundant storage.
  - L-DEO: field data
  - UTIG: processed data
  - NGDC: long-term archive for L-DEO and UTIG data
- USGS/BOEM has a similar approach with NAMSS.
- Needs access to PIs and their students before grants end, to archive field data and products (education/outreach/incentives).

Technical information/issues/challenges:

A. *Desired tools, databases, etc. needed for pursuing key science questions:*

- Continuing access to seismic data acquisition of some kind (see earlier action items).
- Continued integration of existing (and perhaps new) databases.