A Workflows Roadmap for the Geosciences

NSF EarthCube Workflows Community Group

September 15, 2012

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

This work was supported through National Science Foundation under grant # EAR-1238216 as part of the NSF EarthCube initiative.

Preface

The EarthCube Workflows Community Group was formed in March 2012 as part of the NSF EarthCube initiative in response to initial discussions in EarthCube that occurred during 2011. Workflows are used to manage complex computations that have many steps or use large data. Workflow systems assist scientists to select models appropriate for their data, configure them with appropriate parameters, and execute them efficiently. The EarthCube community saw great value in workflow technologies for the future of geosciences.

The goal of the EarthCube Workflows Community Group was to begin to elicit requirements for workflows in geosciences, ascertain the state of the art and current practices, identify current gaps in both the use of and capabilities of current workflow systems in the earth sciences through use case studies, and identify grand challenges for the next decade along with the possible paths to addressing those challenges.

The group was asked to produce a roadmap for workflows in geosciences. Three other Community Groups were formed (Data, Semantics and Ontologies, and Governance), and each was asked to create a roadmap in their area. NSF guidance for the roadmap was to structure it in the following ten sections:

1. Purpose: Introduction, including community(ies) to be served, technical area(s) of the roadmap, and brief discussion what improvements in the present state-of-the-art in geoscience data discovery, management, access, or utilization it will enable. Also include examples of how the outcomes from your effort will enable the community to be more productive and capable.
2. Communication: Description of a communications plan with end users, developers, and sponsors, as well as links to and feedback from other EarthCube community groups and EarthCube concept projects to promote systems integration and accelerate development. Include a discussion of needed interactions with allied fields, agencies, and other related activities (present and desired).
3. Challenges: Description of major drivers, trends, and shifts impacting or that could impact the focus of a working group, including but not limited to changing technology, adoption culture, and community engagement.
4. Requirements: Process(es) to be used to get the necessary technical, conceptual, and/or community (i.e., end-user) requirements at the outset and during the life of the activity, including approaches to achieving community/end-user consensus.
5. Status: Description of the state of the art within the topical area of your roadmap. This should include approaches and technologies from geoscience, cyberinfrastructure, and other fields, the public or commercial sector, etc. that have the potential to benefit the EarthCube enterprise.
6. Solutions: Process for the identification and comparison (pros and cons) of approaches and technology solutions that will contribute to the EarthCube

goal of satisfying current and future research needs of the geoscience end-user.

7. Process: Process(es) to develop community standards, protocols, test data, use cases, etc. that are necessary to mature the functionality of the topical area and promote interoperability and integration between elements of EarthCube.
8. Timeline: Timeline for the project and all related sub-projects, including prioritization of activities and measurable milestones/major achievements and total resources (human and financial) required to achieve roadmap goals over a period of the next 3 to 5 years.
9. Management: Management/governance/coordination plan and decision-making processes necessary to successfully establish standing committee(s) and subcommittees (if warranted), including a plan to identify and respond to shifts in technologies and changing needs at the end-point of use. Include discussion of approaches to educating end-users and achieving community consensus on advancing the capability/technological solution.
10. Risks: Identification of risks and additional challenges to the successful establishment of any working group, and any unique risks associated with a working group associated with your topical area. With respect to identified risks, an approach to risk mitigation should be addressed

The group held a series of virtual and face-to-face workshops to solicit participation from the geosciences community and other relevant researchers.

The EarthCube Workflows Community Group set up a public web site where all their activities were made open for participation from the community and all documents were posted for public access and editing (https://sites.google.com/site/earthcubeworkflow/). Presentations and discussions were recorded and posted on the site.

A key result of the work of the EarthCube Workflows Community Group activities in Spring and Summer 2012 was the creation of a workflows roadmap for the geosciences. An initial roadmap document for the EarthCube community that was first released in June 2012 and presented to the EarthCube community. A revised roadmap was delivered to the community in August 2012. The roadmap serves as a living document created as a group effort with provisions and a process to update and extend it over time.

This document represents the final roadmap of the NSF EarthCube Community Group for workflows in the geosciences. Community feedback is always welcome, as the roadmap will be revised and extended while EarthCube activities continue.

The EarthCube Workflows Community Group

Acknowledgements

This work was supported through National Science Foundation under grant # EAR-1238216 as part of the NSF EarthCube initiative. EarthCube is an innovative and long-term cross-directorate initiative of the US National Science Foundation. The goal of EarthCube is to transform the conduct of research by supporting the development of community-guided cyberinfrastructure to integrate data and information for knowledge management across the Geosciences.

This report was authored by the EarthCube Workflows Community Group. The Steering Committee edited the document, and was chaired by Yolanda Gil (University of Southern California Information Sciences Institute) with the following members: Ewa Deelman (University of Southern California Information Sciences Institute), Ibrahim Demir (University of Iowa) Christopher Duffy (Pennsylvania State University), Suresh Marru (Indiana University), Marlon Pierce (Indiana University), and Gerry Wiener (National Center for Atmospheric Research). The report was assembled from many contributions from the community, notably from Ilkay Altintas (University of California San Diego), Aaron Braeckel (National Center for Atmospheric Research), Steven F. Browdy (OMS Tech, Inc.), David Fulker (OPENDaP), Nicong Li (University of Redlands), Paul Ramirez (NASA/JPL), David Tarboton (Utah State University), and Shahani Weerawarana (Indiana University).

Steering Committee of the Workflows Community Group

Aaron Braekel provides technical leadership within NCAR for the weather component of the FAA NextGen modernization program, in close collaboration with the NOAA/National Weather Service 4-D Weather Data Cube. These real-time operational systems exchange information using standardized web services from the Open Geospatial Consortium (OGC). NCAR's role in the program includes initial research, operational prototyping, system architecture and design, improvements to standards within the relevant bodies, data format development, federated catalogs/registries with standardized metadata, visualization, and dealing with large gridded data volumes. Much of the NCAR's work for FAA was directly adopted for implementation in the NOAA 4-D Wx Data Cube system. Arron also provides technical leadership for development of the weather conceptual model (UML) and exchange model (XML) in collaboration with Eurocontrol. WXXM is currently being considered for worldwide adoption by the International Civil Aviation Organization (ICAO) and the World Meteorological Organization (WMO). Braekel is lead developer for the Java dataserver portion of the Aviation Digital Data Service (ADDS), a website serving general aviation and pilots. The dataserver is a web service exposed on the ADDS websites that written in Java to access a relational database of decoded aviation weather products. The experimental and operational versions of the ADDS websites currently serve around twelve million total hits a day, and the dataserver component accounts for 1 to 1.5 million of those hits.

Dr. Ewa Deelman (http://www.isi.edu/~deelman) is a Research Associate Professor in the Computer Science Department at USC and leads the Collaborative Computing Group at USC/ISI. Dr. Deelman's research focuses on distributed systems with particular emphasis on scientific workflow systems and their application in diverse scientific domains. She is the PI of the NSF-funded Pegasus Workflow Management System project. Pegasus is a collaboration between USC and the Condor group at UW Madison. As part of the Pegasus work, Dr. Deelman developed a number of workflow optimization techniques geared towards improving workflow performance, scalability, and reliability. Pegasus is being used in a number of scientific disciplines including astronomy, bioinformatics, biology, earthquake science, helioseismology, physics, etc. For example Pegasus has been used since 2005 by the Southern California Earthquake Center to run large-scale wave propagation simulations on the national cyberinfrastructure.

Dr. Ibrahim Demir (http://myweb.uiowa.edu/demir/) is an Assistant Research Engineer at the IIHR – Hydroscience and Engineering, University of Iowa. He received his doctoral degree in Environmental Informatics at University of Georgia. His research interests are environmental information systems, scientific visualization of geo-spatial data, user interface design, and information communication. His research efforts are aimed at developing novel information interfaces, using state-of-art visualization techniques, and providing support to researchers from various disciplines to help them visualize and understand environmental observations and simulation data. He has developed interactive visualizations of large, geo-spatial scientific data sets to improve communicating flood related observations and simulation results with public and decision makers. He worked on the design and development of the Iowa Flood Information System (IFIS), a one-stop web-platform to access community-based flood conditions, forecasts, visualizations, inundation maps and flood-related data, information, and applications. Dr. Demir is the member of CUAHSI Informatics Standing Committee.

CUAHSI (Consortium of Universities for the Advancement of Hydrologic Science, Inc.) is an NSF sponsored consortium of 125 universities providing support for the study of the terrestrial components and processes of the global water cycle.

Dr. Christopher J. Duffy (http://web.me.com/cxd111) is a Professor of Civil and Environmental Engineering at The Pennsylvania State Univeristy. Dr. Duffy was a Co-PI with the NSF Science Technology Center on Sustainability of Water Resources in Semi-Arid Regions, which developed new strategies for multi-process hydrologic modeling in river basins of the southwest US. The work lead to development of the Penn State Integrated Hydrologic Model (PIHM). PIHM (http://www.pihm.psu.edu/) is a multiprocess, multi-scale model where the processes are fully coupled using the semi-discrete finite volume method. Duffy is PI of the NSF Critical Zone Observatory at Penn State (2008-2013) where geoscience research focuses on predicting the geochemical, hydrologic, biologic, and geomorphologic processes within the Earth's Critical Zone (http://www.czo.psu.edu/). His recent research focuses on developing a new generation of computational models and supporting data infrastructure for watershed and river basin water and water quality resources. This research involves development of a national strategy to provide researchers, educators and resource managers with seamless and fast access to essential geo-spatial/geo-temporal data, physics-based, high resolution numerical models, and data fusion tools that are necessary to predict and manage the nations surface, groundwater and ecological resources.

Dr. Yolanda Gil (Chair) (http://www.isi.edu/~gil) is Director of Knowledge Technologies and Associate Division Director at the Information Sciences Institute of the University of Southern California, and Research Professor in the Computer Science Department. She received her M.S. and Ph. D. degrees in Computer Science from Carnegie Mellon University. Dr. Gil leads a group that conducts research on various aspects of Interactive Knowledge Capture. Her research interests include intelligent user interfaces, knowledge-rich problem solving, discovery informatics, scientific workflows, and the semantic web. She has developed novel representations and algorithms for semantic workflows, which can reason about constraints of data and software components within the workflow. Dr. Gil and her colleagues developed the Wings workflow system (http://wings.isi.edu) and a number of workflow applications for earthquake simulations, ecology, genomics, and more recently drug discovery. Dr. Gil was elected to the Council of the American Association of Artificial Intelligence (AAAI), and served in the Advisory Committee of the Computer Science and Engineering Directorate of the National Science Foundation. She recently chaired the W3C Provenance Group, an effort to chart the state-of-the-art and posit standardization efforts in this area. In 2010 she was elected Chair of ACM SIGART, the Association for Computing Machinery's Special Interest Group on Artificial Intelligence.

Suresh Marru (http://people.apache.org/~smarru/) directs the Science Gateway efforts of the NSF funded Extreme Science and Engineering Discovery Environment (XSEDE). Marru has previously participated in geoscience efforts including Linked Environments for Atmospheric Discovery (LEAD), Coupled Modelling Environment of Atmospheric Discovery (MEAD) Expedition, and Environment Hydrology Application Teams (EHAT) Alliance projects. These projects are few of the many efforts which have set the stage for EarthCube. As one of the co-principal investigators and architects of the LEAD workflow infrastructure, Marru was instrumental in creating an integrated, scalable framework in which meteorological analysis tools, forecast models, and data repositories can operate as dynamically adaptive, on-demand, grid-enabled systems. Marru's current research focus is to work alongside multidisciplinary experts in the field of eScience to developed open community driven web and service-based scientific workflow and gateway systems for individual research, collaboration, outreach, and cross-disciplinary collaboration. Marru and Pierce are the founding members of the open community driven scientific workflow framework Apache Airavata.

Dr. Marlon Pierce (http://pti.iu.edu/sgg) is the leader of the Science Gateways Group in the Research Technologies division of University Information Technology Services at Indiana University. Pierce received his M. S. and Ph. D. degrees in Physics from Florida State University. Pierce's team conducts research and development in the application of distributed computing and information technologies to problems in applied fields such as earthquake science, astronomy, astrophysics, computational chemistry, material science, atmospheric science, and bioinformatics. Pierce's Science Gateway Group is also actively involved in open source software development and governance, acting as founding members of the Apache Rave and Apache Airavata incubator projects. The focus of this work is to bring open community software engineering methodologies to cyberinfrastructure software development. Pierce's service to the cyberinfrastructure community includes the organization of the Gateway Computing Environments (GCE) workshops, part of the annual ACM/IEEE Supercomputing conference. GCE workshops provide a venue for peer reviewed publications on the state of the art in Science Gateway development.

Dr. Gerry Wiener is a senior engineer at the Research Applications Laboratory (RAL) at the National Center for Atmospheric Research. He has been responsible for meteorological system design and implementation at NCAR since 1987 and is the engineering deputy of the Weather Systems and Assessment Program at RAL. Dr. Wiener is currently the lead software engineer on the Xcel Wind Energy project at NCAR/RAL. He oversaw the design and implementation of a wind/power forecasting system that makes hourly operational wind energy forecasts out to 168 hours for all Xcel Energy wind farms in Colorado, New Mexico, Texas and Minnesota. He has been the lead software engineer on the Graphical Turbulence Guidance (GTG) project for the FAA for over 10 years and successfully transferred multiple versions of the Graphical Turbulence Guidance (GTG) system to the Aviation Weather Center in Kansas City. The GTG output grids are now an official National Weather Service product displayed on the Aviation Digital Data Service. Dr. Wiener was also the lead software engineer on the Wind shear and Turbulence Warning System for the Hong Kong Government. This system has been used operationally at the Hong Kong International Airport for detecting hazardous turbulence and wind shear since July 1998. Dr. Wiener is interested in finding ways to build meteorological processing systems using Workflows, Data Mining and Data Cube technology.

Executive Summary

Introduction

Advances in geoscience research and discovery are fundamentally tied to data and computation, although formal strategies for managing the diversity of models and data resources in the earth sciences have not yet been resolved or fully appreciated. Through this roadmap document we hope to motivate the importance of scientific workflows in support of geoscience research, and discuss a comprehensive path towards achieving the goals and practical benefits of leveraging scientific workflows in geoscience research and discovery.

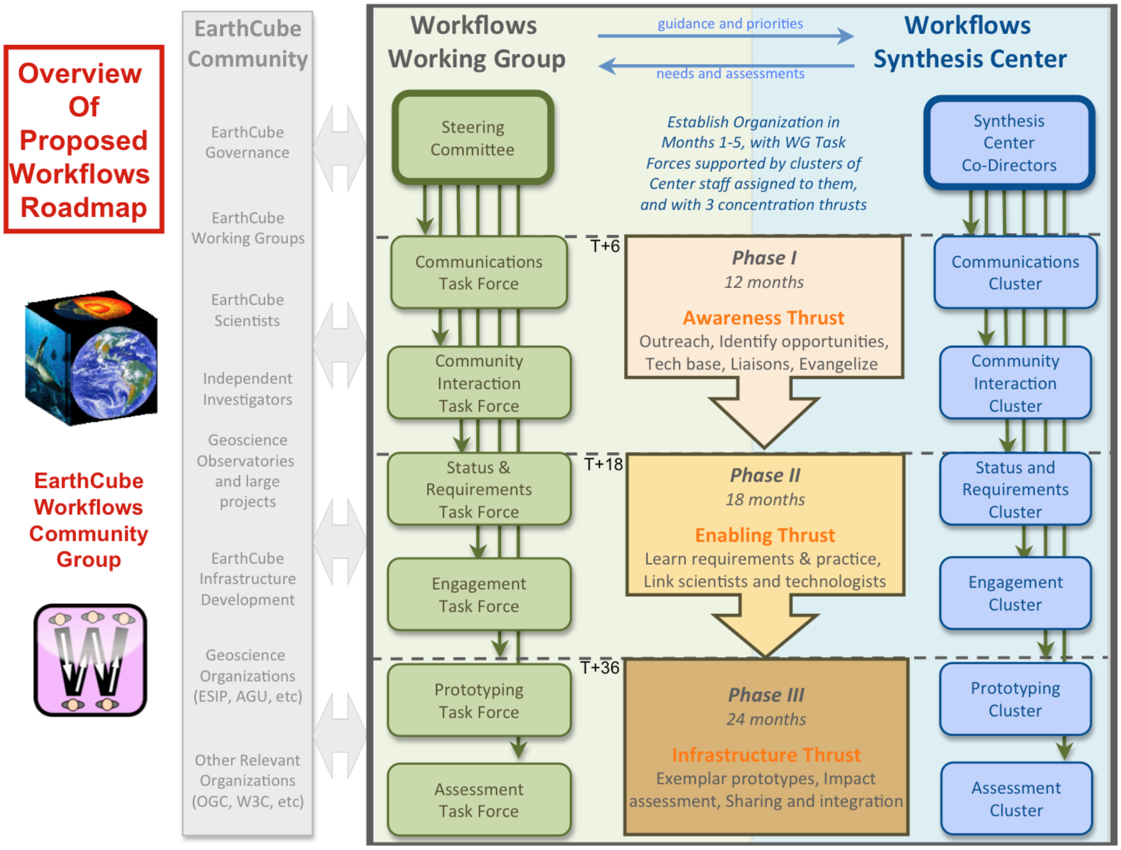
Scientific activities can be seen as collections of interdependent steps represented as *workflows*. Gathering and analyzing data, coordinating computational experiments, and publishing results and data products are organized activities traditionally captured in research notebooks. Today we have the ability to digitally codify much of these activities, particularly for computational experiments, using workflow technologies. Workflows may be used to execute enormous computations, to combine distributed data and computing resources in novel ways, and to guide scientists through complex processes. When combined with metadata and provenance-capturing capabilities, workflows allow reproducibility of results, increased efficiency, and enhanced publications. The challenge before us is to make these tools ubiquitously available, enhanced, and adopted for the geosciences.

The EarthCube Workflows Community Group was created as part of the NSF EarthCube initiative. Its goal is to constitute a broad community within the geosciences that will identify both short-term problems and long-term challenges for scientific workflows. Addressing this goal is the central theme of this roadmap. Aspects of this goal include better education and outreach, better understanding of the different types of workflows, better collaboration between workflow software developers and geoscientists, the identification of gaps, and the vision for grand challenges that no workflow technology can currently address. The resulting workflow roadmap is considered a living document that will be extended and updated as future needs and our understanding of the problems evolve.

A cornerstone of the roadmap is to bring to life a **Workflows Synthesis Center** that will provide resources for an EarthCube **Workflows Working Group**, providing together an umbrella for all workflow-related Earthcube activities and for coordination with other activities that focus on other aspects of EarthCube. Specific task forces are identified in this roadmap.

The Workflow Working Group's Steering Committee will be the central organizer and broker with the EarthCube community. In its initial phase, the Steering Committee members were invited by the NSF to bootstrap the Workflows Group. In the next phase, this Steering Committee must expand to include new members, both funded and unfunded, who will be stakeholders in the working group. The Steering Committee will thus need to include representative end users, workflow researchers, representatives of other relevant projects, liaisons with other EarthCube working groups, and funding agency representatives.

A graphical overview of the Workflows Community Group Roadmap is shown below. The overall timeline is highlighted vertically in the middle, with the Workflows Working Group and the Workflows Synthesis Center interacting synergistically to support the roadmap activities.


Overview Of Proposed Workflows Roadmap
EarthCube Workflows Community Group
EarthCube Community
EarthCube Governance
EarthCube Working Groups
EarthCube Scientists
Independent Investigators
Geoscience Observatories and large projects
EarthCube Infrastructure Development
Geoscience Organizations (ESIP, AGU, etc)
Other Relevant Organizations (OGC, W3C, etc)
Workflows Working Group
guidance and priorities
needs and assessments
Workflows Synthesis Center
Steering Committee
Establish Organization in Months 1-5, with WG Task Forces supported by clusters of Center staff assigned to them, and with 3 concentration thrusts
Synthesis Center Co-Directors
T+6
Communications Task Force
Phase I
12 months
Awareness Thrust
Outreach, Identify opportunities, Tech base, Liaisons, Evangelize
Communications Cluster
Community Interaction Task Force
Community Interaction Cluster
T+18
Status & Requirements Task Force
Phase II
18 months
Enabling Thrust
Learn requirements & practice, Link scientists and technologists
Status and Requirements Cluster
Engagement Task Force
Engagement Cluster
T+36
Prototyping Task Force
Phase III
24 months
Infrastructure Thrust
Exemplar prototypes, Impact assessment, Sharing and integration
Prototyping Cluster
Assessment Task Force
Assessment Cluster

Communications

Effective communication plans and mechanisms will be essential for meeting the workflow roadmap goal for ubiquitous adoption of workflow technologies. Current communication shortcomings include lack of awareness of workflow technologies by geoscientists and lack of understanding of geoscience requirements by workflow researchers and developers. This leads to problems such as invention of redundant, individually unsustainable tools and lost opportunities to collaborate on long-term challenges such as scientific reproducibility and operational efficiency. To be successful, we must address the communication barriers between geoscientists and cyberinfrastructure researchers.

A Communications Task Force will create connections with the community, materials for dissemination of workflow concepts and opportunities, and engagement activities such as workshops and virtual meetings.

Challenges

The overall goal of the Workflow Roadmap will be to make workflows ubiquitous within the geosciences and to further develop or enhance the workflow tools to meet the needs of geosciences. Several challenges must be overcome to reach this goal.

Technical Challenges: Workflow researchers are constantly gathering requirements from the scientific community, which are sophisticated and beyond reach of current technologies. Basic research needs to be done in the context of EarthCube requirements, as the capabilities required to support the EarthCube vision only exist in part. The group will have to develop mechanisms to facilitate early transition of new capabilities to users. To ensure success, these activities need to occur as a partnership between scientists and developers as new workflow capabilities are addressed.

Broader Adoption: While there are a number of workflow systems that are used and/ or well-known in the geosciences community, there is also much reinvention and lack of use. The tension between encouraging adoption of mature workflow systems versus development of lightweight customized systems or simple scripting solutions will need to be addressed. A large percentage of geoscientists are not using any workflow tool. This has a number of consequences: lost efficiency, lost metadata, lack of reproducibility, limited or no access to national geoscience datasets, problem of national geo-spatial/temporal data on secure federal servers with many different formats, etc. The challenge is to increase the access and efficiency of access of workflow technologies to geoscientists.

Reproducibility: Reproducibility, a cornerstone of the scientific method was identified as an important problem in interactions to date with the community. Reproducibility will require using semantic representations that document enough details about scientific processes in a reusable form, so they can be easily re-run by others and adapted to new problems. True bit-by-bit reproducibility may be an impossible problem as heterogeneous execution platforms may generate slightly different results. However, the more coarser reproducibility--the scientific reproducibility needs to be attained.

Rapidly Evolving Technologies: Resources available to scientists are changing rapidly, challenging cyberinfrastructure (and particularly workflows) to stay in step. While evolving infrastructure enables more powerful computations and the storage of more data, it also introduces impedances to integration, such as the difficulties of moving data, provisioning adequate storage, computational resources, dealing with various security mechanisms, etc.

Requirements

We will need an ongoing process for obtaining, understanding and evaluating the requirements of the geoscientific community. The diversity of users is an important challenge that must be addressed when obtaining these requirements. Our first step was to outline typical use-case workflows that form the organizing principal of the Workflow Roadmap. As part of its March-June 2012 workshop series, our next step was to create a the Workflows Community Group questionnaire as a way to capture community input. The survey format allowed essay responses to questions. From the community survey responses so far obtained, efficient sharing of multi-step data transformations, handling big data, projecting diverse geospatial/temporal data sets, integrating multiple data sets, managing complex executions, reproducibility of results, and interoperability with other tools and services (OPENDaP, NetCDF, OGC services, ArcGIS, etc.) are all capabilities mentioned by the responders. The responders covered a full range of geoscience research.

The next step will be to begin to develop/design a basic strategy for aligning broad user requirements determined from the surveys and workshops with existing and novel workflow technologies. We will need to develop a matrix associating use-cases with workflow technologies that recognize the particular data and model needs in each case, that allow for automated management and sharing of information, integrate resource planning and scheduling, data quality assurance and generally provides a test-drive of a new vehicle for research discovery.

A Prototyping Task Force will test the technical requirements posed by the community with prototypes of typical use-cases. The process will require follow-up, evaluation, and community consensus on all phases of the Workflow Roadmap.

Status

In general, it is difficult to assess the current state of the art in the various fields and commercial sectors. This type of assessment needs to happen as part of an ongoing earthcube activity (both because of the scope of the activity and the dynamic nature of the state-of-the art technologies). This activity can be done by the Status and Requirements Task Force through interactions with the science-focused workshops planned for the Fall of 2012. The roadmap surveys workflow solutions from the geoscience community, the cyberinfrastructure community, and commercial vendors.

Solutions and Process

The Workflows Working Group will include an Engagement Task Force that will: 1) provide guidance to geoscientists in identifying approaches to address their workflow needs, 2) assist scientists in evaluating potential workflow technology solutions, 3) request the support of the Status and Requirements Task Force and the Prototyping Task Force when necessary 4) disseminate their expertise in workflow solution approaches.

The underlying basis for identifying approaches to address workflow needs is the creation of a situation specific workflows capability maturity model. The process to identify technology solutions is based on a technology evaluation framework.

The roadmap outlines processes for identifying workflow approaches and identifying technical solutions. Processes will also be used for identifying appropriate standards, developing use cases and reference implementations through open community processes.

A Community Interaction Task Force will document use cases for existing and potential uses of workflows in the geosciences. It will also identify and prioritize needs for basic research in workflows motivated by grand challenges in the geosciences, and facilitate transfer of new advances in workflows research into geoscience infrastructure and adoption by scientists. It will also document existing standards and recommendations for adoption and interoperability.

An Assessment Task Force will track and assess the impact of workflow technologies across geosciences through the collection of quantitative and qualitative data at the early stages of EarthCube and as the roadmap activities progress.

Timeline

The overarching goal of the workflows working group is to make workflows ubiquitous within the geosciences community. This roadmap is motivated towards addressing the challenges we see in the community as extensively discussed in the previous sections. The biggest issues against achieving the overarching goal of ubiquity is the lack of awareness on how to map science challenges into workflow technologies that would improve the process, and the diverse and dispersed workflow community. Hence our activity prioritization, our milestones and the associated timelines are heavily influenced towards addressing the major issues early on, in the roadmap execution.

The timeline is divided into three overlapping thrusts: 1) Awareness Thrust, focused on community outreach and requirements gathering, 2) Enablement Thrust, focused on prototyping proofs of concept and working with the community to disseminate workflow technologies, 3) Infrastructure and Services Thrust, focused on developing community infrastructure that would include workflow publication and citation, workflow sharing, workflow execution resources, and other substantial community resources concerning workflows.

Management

The goal of management is to execute the roadmap and its principal goal of making workflows ubiquitous within the geosciences.

The working group management must be as effective as possible. Traditional management processes are inadequate for the roadmap evolution and execution, since we must coordinate multiple independent organizations and individuals. Thus the problems that we need to solve can be categorized into two perspectives - organizational and individual.

The challenges from an organizational perspective include establishing an effective organizational structure that would enable efficient strategizing; establishing an effective operational structure that would ensure smooth and timely operational activities; establishing effective processes for creating groups, organizations, etc; and establishing effective processes to facilitate consensus and enable efficient decision-making.

The problems that need to be addressed from an individual's perspective include efficient and productive use of participants' time, creating incentives beyond funding to encourage participation, and supporting and rewarding initiative by individuals;

The organizational goals will primarily be achieved through the organizational structure of the workflows working group and the associated open community process model. The goals from an individual's perspective will to a large extent, be facilitated by the substructures within the overall organizational structure, and the associated open community process model.

The strategy will be to establish a central Steering Committee for the Working Group with the flexibility that allows its members to take initiative to address problems. We also plan on establishing an institute that would function as a Synthesis Center that will support the Working Group, and specific Task Forces that would enable the implementation of the strategies proposed in this roadmap. Each of these structural components and their operational processes and primary responsibilities are discussed in detail within the roadmap.

Risks

A number of risks have been identified, including not establishing meaningful requirements, substantive differences in user requirements, not addressing workflow requirements, inadequate communication with the scientific user community, lack of adoption, and choosing the wrong software engineering methodology. The primary mitigation mechanism is the implementation of a community-based governance model that will be specifically charged with representing the community.

Section 1. Purpose

This section includes an introduction, describes community(ies) to be served, technical area(s) of the roadmap, and brief discussion what improvements in the present state-of-the-art in geoscience data discovery, management, access, or utilization it will enable. It also includes examples of how the outcomes from the workshops will enable the community to be more productive and capable.

1.1 Introduction

Advances in geoscience research and discovery are fundamentally tied to data and computation, although formal strategies for managing the diversity of models and data resources in the earth sciences have not yet been resolved or fully appreciated. Through this roadmap document we hope to illustrate the importance of scientific workflows in support of geoscience research, and discuss a comprehensive path towards achieving the goals and practical benefits of leveraging scientific workflows in geoscience research and discovery.

Scientific activities are workflows. Gathering and analyzing data, coordinating computational experiments, and publishing results and data products are organized activities traditionally captured in research notebooks. Today we have the ability to digitally codify much of these activities, particularly for computational experiments, using workflow technologies. Workflows may be used to execute enormous computations, to combine distributed data and computing resources in novel ways, and to guide scientists through complex processes. When combined with metadata and provenance-capturing capabilities, workflows allow reproducibility of results, increased efficiency, and enhanced publications.

The goal for the EarthCube Workflow Working Group is to make workflow tools an everyday part of every geoscientist's research. Addressing this goal is the central theme of this roadmap. Aspects of this goal include better education and outreach, better understanding of the different types of workflows, better collaboration between workflow software developers and geoscientists, the identification of gaps, and the identification of grand challenges that no workflow technology can currently address.

The concept of a *workflow* is simply a description of the processes by which geoscientists work with data, models, information, as well as their students and colleagues to contribute to Earth Science knowledge (Deelman et al, 2005). The workflows enabled by the emerging data and information systems have the potential to bring together data from a variety of disciplines that can be easily processed to fill gaps in information and knowledge and facilitate the translation of the new science in action with the participation of both scientists and practitioners. These systems can also improve the information services and operational science needed for efficient and effective decision making. The information services provided by these systems allow the establishment of an iterative dialogue between the information users and producers through and open and transparent cyber platform accessible to all involved stakeholders (i.e., scientists, emergency planners, natural resource managers, policy makers, and the public at large).

The EarthCube Workflows Community Group was created as part of the NSF EarthCube initiative. Its goal is to constitute a broad community within the geosciences that will identify both short-term problems and long-term challenges for scientific workflows. The resulting workflow roadmap will be considered a living document that will be extended and updated as future needs evolve.

1.2 Scientific Workflow(s)

This section gives a brief introductory overview about what scientific workflows are and their role in science, research, and practice.

1.2.1. Workflows in General

In general, a workflow is a series of connected steps employed to accomplish some overall goal. In this simplest cases, each step produces output that the next step needs as input and the steps are repeated every time new input is available at the first step. Almost all geoscientists employ some kind of workflow in their work while processing data from a sensor grid, cataloging data from the field, visualizing data, or analyzing output from a numerical model. They might invoke a series of applications which each add to or transform data. In most cases, these workflows are implemented informally: the scientists remember (and log or document) which steps to take next. Other workflows may be implemented in a shell script and provided by an institution for others to use or used internally before providing final data products. When workflows, even simple ones, can be recorded in some standard way, the workflow descriptions become powerful data on their own for advancing science.

1.2.2. Workflows as a discipline

A workflow may comprise thousands of steps, where each step may integrate data sources and diverse models developed by different groups. The applications and data may be also distributed in the execution environment. The assembly and management of such complex distributed computations present many challenges, and increasingly ambitious scientific inquiry is continuously pushing the limits of current technology. “Workflows” have recently emerged as a paradigm for representing and managing complex distributed scientific computations and therefore accelerate the pace of scientific progress.

1.2.3. Introductory Materials about Workflows

Workflow related terms and definitions are available as “Workflow Glossary” in the “About Workflows” section of the EarthCube Workflow Community Group site (https://sites.google.com/site/earthcubeworkflow/about-workflows).

There are a number of introductory books and survey articles about workflows, including:

- "Workflows for e-Science: Scientific Workflows for Grids." Ian J. Taylor, Ewa Deelman, Dennis B. Gannon, and Matthew Shields (Eds). Springer Verlag, 2006. Available at Springer.
- "Examining the Challenges of Scientific Workflows." Yolanda Gil., Ewa Deelman, Mark Ellisman, Thomas Fahringer, Geoffrey Fox, Dennis Gannon, Carole Goble, Miron Livny, Luc Moreau, and Jim Myers. IEEE Computer, 40(12):24-32, 2007. Also available at computer.org and here as a free pre-print.
- “From Data to Knowledge to Discoveries: Artificial Intelligence and Scientific Workflows.” Yolanda Gil. *Scientific Programming*, 17(3):231-246, 2009. Also available at Meta Press and here as a free pre-print.

1.3 Communities To Be Served

The Workflow Community Group actively seeks participation from the geosciences, cyberinfrastructure, and other relevant communities that can contribute to the formulation of this roadmap. We will coordinate with the three other newly formed groups. The Roadmap for Geoscience Workflows will be coordinated with the roadmaps of other relevant community activities, which include Semantics and Ontologies, Data Discovery, Mining and Integration, and Governance. Semantics and Ontologies addresses mechanisms for describing workflows and the data they process. Data discovery, mining and integration processes are closely tied to the processing and provenance of particular datasets. Governance activities are important for determining the scientific community’s priorities as well as processes for workflow software development.

Numerous existing groups and federal agencies are potential stakeholders in EarthCube workflow activities. Section 2 describes our activities to date to establish communication lines with specific communities and plans for future activities in this area. To illustrate how workflows might impact scientists and communities in geosciences, we have developed a handful of prototypical “Workflow Use Paradigms”. We also describe “vignettes” of current or potential uses of workflows in the community, to serve as illustrations of the benefits of using workflow technologies.

1.3.1. What are the Common Workflow Use Paradigms?

Here are common workflow use paradigms, or profiles, capture modes of use of workflow technologies. For scientists that do not use workflow technologies, these paradigms can provide models of reference and a better understanding of how workflow technologies can be relevant to their work and help improve the state-of-the-art enabled by this roadmap.

i. The Center Modeling Paradigm: A group of people (a Center) design the workflow to run specific models, the results of executing the workflow are specialized data products that are then published to a community of users.

An example of this is the Iowa Flood Center (although they don't currently use a workflow system). A group of scientists develop the models and run them, they integrate data from a specific set of sensors. Once the workflows are run then the results are published so that all stakeholders in the Center can benefit. In the future, other people may want to try the Center Modeling Workflow but with of their own models, and adapt the workflow with their own data. This might be combined with the ***The Social Paradigm*** *workflow below.*

ii. The Long Tail Paradigm: An individual investigator runs her own workflow, she prepares the data herself, integrates it with data from shared sources, and runs her own models.

An example of this is a river ecologist, interested in estimating water metabolism rates. She sets up her own local sensors, prepares the data, integrates it with data from government sites/services (e.g. weather, other sensors in the area, etc), and then runs some models. Currently this investigator might just use spreadsheets to prepare the data and shell scripts to manage the model codes, keeping written notes to log what has been done. In the future, they would like better support to do the routine data preparation, so sensor data is pre-processed automatically, and they would like to easily incorporate into their workflow the models developed by people, run their workflow on other people's data, etc. One issue for the Long Tail Paradigm is that the individual or small team are often place-based will likely still require "National Data Sets" as a basis for their workflow. In the ecologist example they might need what we can call Essential Terrestrial Data" such as Geospatial: topography (DEM), land cover, land use, soils, geology, weather, climate, etc. Geotemporal or point data: streamflow, water quality, sediment load, stream temperature, etc. All Geospatial / Temporal data would need to have consistent formats to be useful. The scientist may also publish her data and workflows in an open repository for others to use it.

iii. The High Performance Computing Paradigm: Very large data-sets are analyzed or produced with computationally intensive codes, requiring high-end computing resources and scalable infrastructure.

An example of this is running climate models and analyzing their output. A subset of the Big Data Paradigm is the ***Data Intensive Paradigm*** *where large computational models require very large data-sets during runtime as well as pre- post-processing. An example of this is Ecological and Hydrologic models using vegetation, land use, digital terrain, soils, weather and climate data as input to their models. The workflow requires usual HPC for the models, but also requires handling large data during run-time.*

iv. The Whiteboard Paradigm: In this example a group of investigators in a collaboration on an interdisciplinary investigation drafts a sketch of what looks like a workflow, just to gain an understanding on what models they are agreeing to run on their individual data. The team may have little understanding of each others needs at first but want to work together to solve some big science or multidisciplinary science question.

> *An example of this is lake researchers (eg GLEON). A generalization of this might be called* ***The Social Paradigm*** *or* ***The Interdisciplinary Paradigm*** *where a group of investigators in an interdisciplinary investigation has little understanding of each others needs at first but want to work as a team to solve some big science or multidisciplinary science question. An example of* ***The Social Paradigm*** *might be ecologists, hydrologists, coastal marine and social scientists working on the Chesapeake Bay watershed-estuary problems or possibly the Critical Zone Observatory program where geologists, hydrologists, geo-chemists, ecologists, weather, climate and soil scientists are all working together and want to use each others science in the most productive way. The important point here is the workflow needs a way to support interdisciplinary teams trying to resolve very hard holistic problems or trying to solve very specialized science with a multidisciplinary worldview of data and models.*

v. The Metadata-Rich Paradigm: A collaboration that wants to track metadata of all data used and produced by their analyses. The user in the Metadata-Rich Paradigm may be individuals or communities of geoscientists where data collection is a very intensive process and often uses very sophisticated laboratory analyses on each sample they collect.

> *An example of this is geochemists who collect limited amounts of actual rock sample data for all rocks of the earth, but processing and lab analyses lead to very large amounts of metadata.*

vi. The Software Marketplace Paradigm: Scientists in a discipline often use common scientific methods and best practices. Capturing these methods and best practices as workflows, and facilitating their inspection and reuse would be beneficial for many science communities.

> *Consider the data preparation steps that need to be done to put data in a format appropriate for a model. These steps may include simple reformatting, but also other transformations that address quality control. In many disciplines, these kinds of workflows are often redundantly rebuilt by different groups who use data from the same shared repositories or for the same models. Data preparation is estimated to take 60% of the effort in science projects. Workflow community repositories would greatly reduce this effort.*

These different workflow use paradigms are summarized in the table below.

Workflow Use Paradigm	***Description***
Center Modeling Paradigm	A group of people (a Center) design the workflow to run specific models, the results of executing the workflow are specialized data products that are then published to a community of users.
Long Tail Paradigm	An individual investigator runs her own workflow, they prepare their data themselves, integrate it with data from shared sources, and run their own models.
High Performance Computing Paradigm	Very large data-sets are analyzed or produced with computationally intensive codes, requiring high-end computing resources and scalable infrastructure.
Whiteboard Paradigm	A group of investigators in a collaboration drafts a sketch of what looks like a workflow, just to gain an understanding on what models they are agreeing to all run on their individual data.
Metadata-Rich Paradigm	A collaboration that wants to track metadata of all data used and produced by their analyses. The user in the Metadata-Rich Paradigm may be individuals or communities of geoscientists where data collection is a very intensive process and often uses very sophisticated laboratory analyses on each sample they collect.
Software Marketplace Paradigm	A community of scientists in a discipline share workflows that capture commonly used data analysis methods that reflect best practices, rminimizing duplication of effort to re-create the methods.

1.3.2 Vignettes from the Geoscience Community with Workflows

The range of workflow requirements and considerations is best illustrated by examples. The workflow group is developing "workflow vignettes" that visually and concisely highlight existing or potential uses of workflows in geosciences. The group welcomes vignette submissions from everyone in the community. Each vignette consists of:

- 2-3 sentences describing the goal of using workflows environment, mentioning who are the users of the workflow and/or its results
- 2-3 sentences describing why workflows are useful (reuse, verification, provenance, etc)
- a graphic, e.g. a workflow sketch or a data product of a workflow
- a sentence mentioning the institutions involved, a point of contact, and a URL if available
- (optional): a list of major steps involved in the workflows

We include here a few of the vignettes that we have collected to date.

A Center Modeling Paradigm Highlight:
Aviation Turbulence System Workflow

This system built for the FAA ingests meteorological numerical model data, pilot reports, in situ turbulence measurements using airline accelerometer data, and radar data. It then integrates these data sets and develops a gridded turbulence forecast grid. This grid is distributed to the aviation user community in both visual form and also using GRIB2 files.

The following web link contains current system output:
http://www.aviationweather.gov/adds/turbulence/turbnav

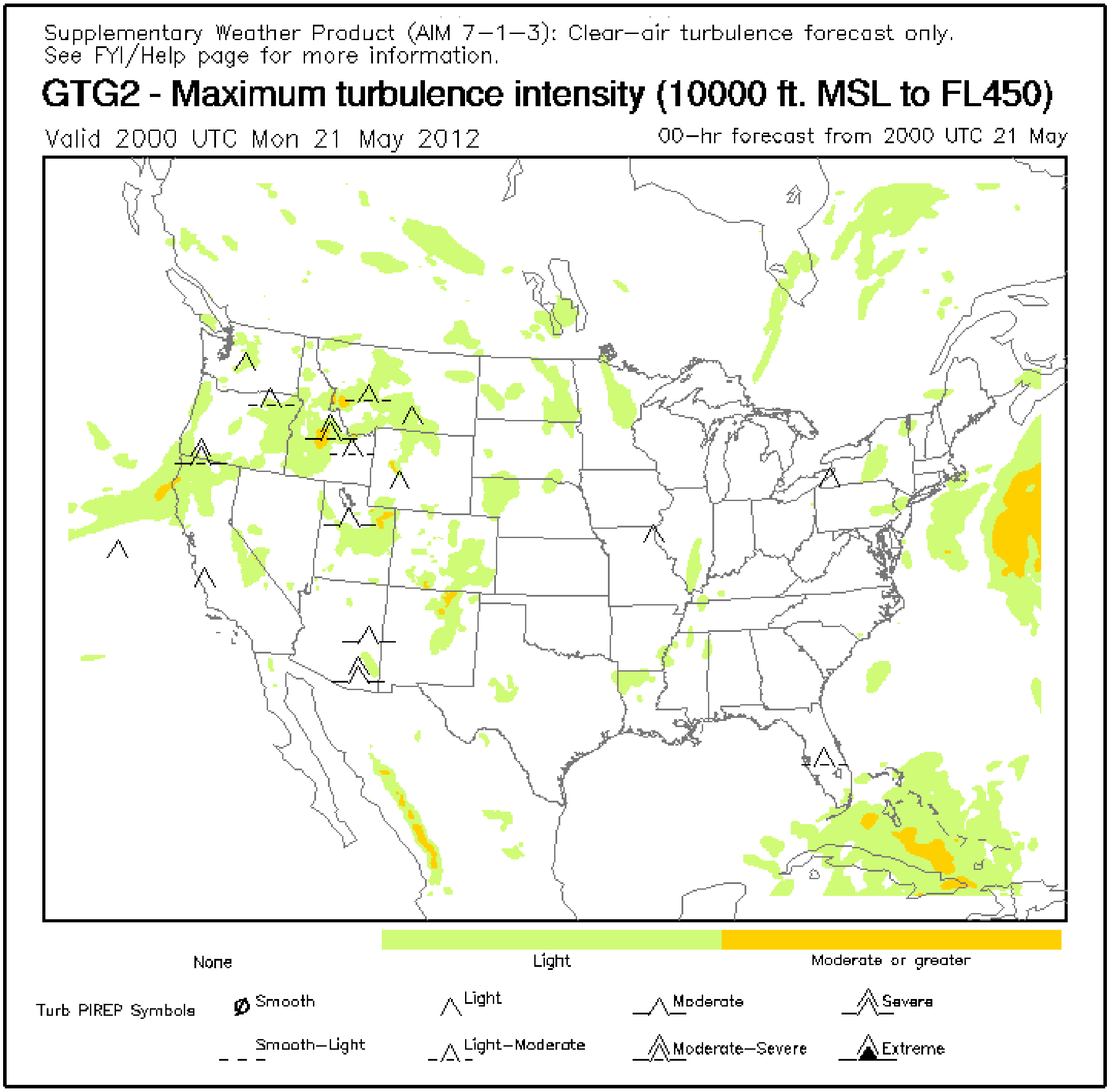
The next link contains additional information with regard to the system:
http://aviationweather.gov/exp/gtg/info.php

This particular system can be categorized within the Center Modeling Paradigm since it runs at the Aviation Weather Center and provides output products to aviation weather users including commercial pilots, the general aviation community and military pilots.

The current system is run using a Python script glue layer that does not incorporate standard workflow technology. The usage of standard workflow technology would potentially benefit ongoing system development, system operations, system support and maintenance.

Workflow Steps

1. Ingest meteorological numerical models such as the WRF or GFS models
2. Ingest pilot reports
3. Ingest in situ turbulence reports
4. Ingest radar data
5. Project all ingest data sets to a common three dimensional grid. In the vertical interpolate to appropriate flight levels at 1000 foot intervals up to 45000 feet
6. Create turbulence diagnosis grids using a variety of turbulence algorithms. The diagnosis grids are based on numerical model input
7. Create turbulence detection grid using radar data
8. Blend turbulence diagnosis grids, turbulence detection radar grid, pilot reports and in situ turbulence reports into turbulence nowcast grid
9. Create turbulence forecast grid by forming a weighted sum of the turbulence diagnosis grids for different forecast lead times
10. Create output forecast files using GRIB2 format
11. Create web images consisting of turbulence forecast grids at different flight levels


Supplementary Weather Product (AIM 7-1-3): Clear-air turbulence forecast only.
See FYI/Help page for more information.
GTG2 - Maximum turbulence intensity (10000 ft. MSL to FL450)
Valid 2000 UTC Mon 21 May 2012
00-hr forecast from 2000 UTC 21 May
None
Light
Moderate or greater
Turb PIREP Symbols
Smooth
Light
Moderate
Severe
Smooth-Light
Light-Moderate
Moderate-Severe
Extreme

A Center Modeling/High Performance Computing Paradigm Highlight: Wind/Power Prediction System Workflow

This system built for Xcel Energy ingests meteorological numerical model data, surface observation data, airplane observations, soundings, turbine wind and power data and transmission node data. It runs a number of numerical models including the WRF RTFDDA and MM5 RTFDDA model ensembles. It incorporates a statistical postprocessing system that integrates all the different numerical model forecasts based on their individual performance with respect to actual wind speed observations. The system then produces forecast winds and power at approximately 100 wind farms and delivers these forecasts and related information to utility system operators and meteorologists at Xcel Energy.

The following web link contains additional information about this system:

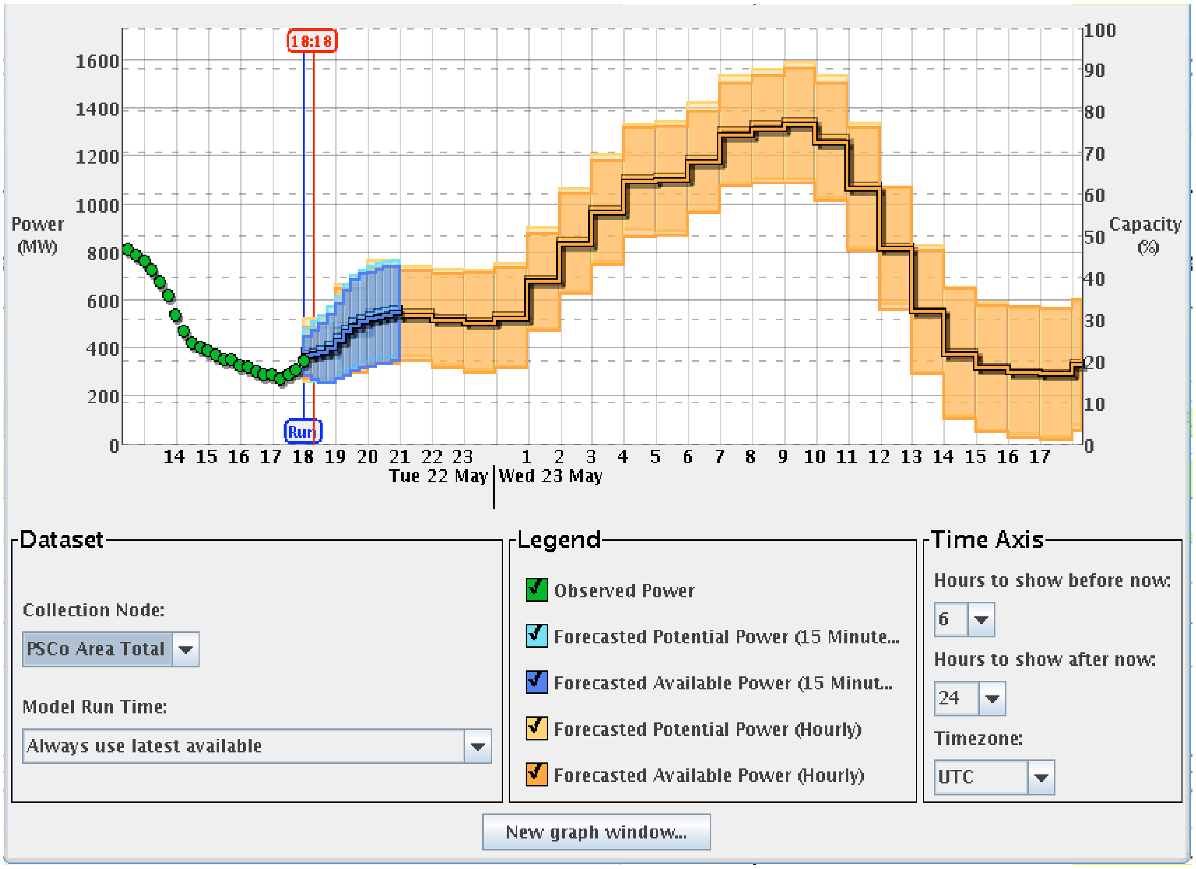
https://www2.ucar.edu/atmosnews/news/5771/ncar-wind-forecasts-save-millions-dollars-xcel-energy

This particular system can be categorized within the Center Modeling Paradigm since it is run by a weather vendor and provides output products to different wind energy producers.

The statistical postprocessing and power conversion portions of the current system are run using a Python script glue layer that does not incorporate standard workflow technology. The usage of standard workflow technology for this portion of the system would potentially benefit ongoing system development, system operations, system support and maintenance.

Workflow Steps

1. Ingest meteorological numerical model data from WRF, GFS, GEM, ECMWF models
2. Ingest surface observation data
3. Ingest airplane observations
4. Ingest soundings
5. Ingest turbine wind, power and connection node data
6. Run WRF RTFDDA model
7. Run MM5 model ensemble
8. Run WRF model ensemble
9. Integrate model output from all models using statistical postprocessing subsystem
10. Create wind forecasts for all wind farms
11. Apply power conversion algorithm to wind forecasts
12. Prepare wind/power forecast output containing wind/power forecast information for each wind farm
13. Update visualization display containing forecast information for all wind farms and regions



Dataset

Collection Node: PSCo Area Total

Model Run Time: Always use latest available

Legend

☑ Observed Power

☑ Forecasted Potential Power (15 Minute...

☑ Forecasted Available Power (15 Minut...

☑ Forecasted Potential Power (Hourly)

☑ Forecasted Available Power (Hourly)

Time Axis

Hours to show before now: 6

Hours to show after now: 24

Timezone: UTC

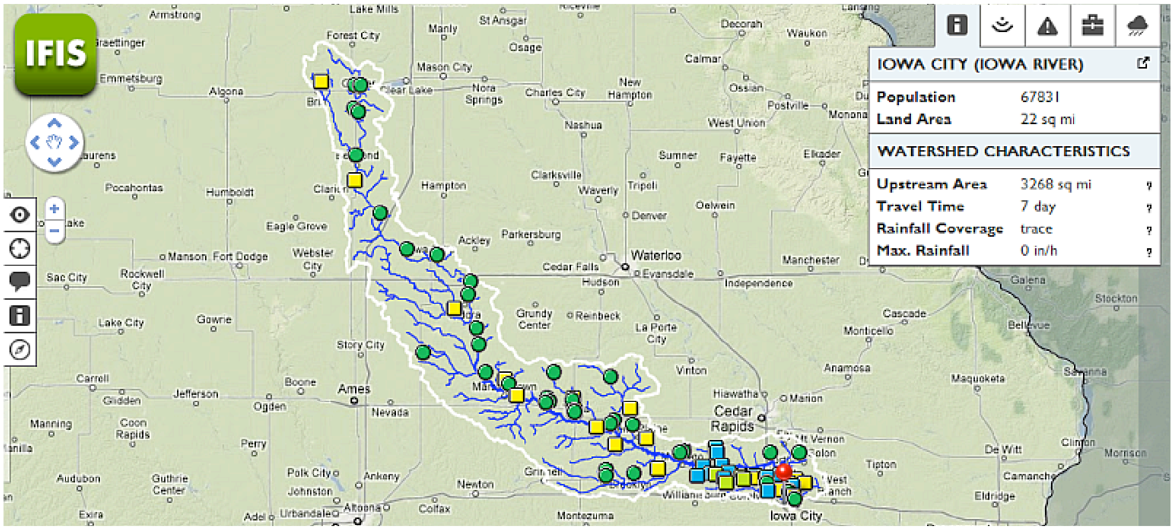
New graph window...

A Center Modeling Paradigm Highlight:
Iowa Flood Information System: Towards Integrated Data Management, Analysis and Visualization

The Iowa Flood Information System (IFIS) is a web-based platform developed to provide access to flood inundation maps, real-time flood conditions, flood forecasts both short-term and seasonal, flood-related data, information and interactive visualizations for communities in Iowa. The IFIS is not currently utilizing a workflow system. It fits the center modeling paradigm since it integrates specific models and shares specialized data products to general public. The IFIS integrated data and models from internal and national resources which requires extensive manual work and expertise. Many steps required to integrate and develop tools for IFIS can be automated with workflow systems. Workflow systems could also provide flexibility to modify the models and visualization tools in real-time. The IFIS helps communities make better-informed decisions on the occurrence of floods, and will alert communities in advance to help minimize damage of floods. More details at http://ifis.iowafloodcenter.org

Workflow Steps

1. Ingest precipitation data products from nexrad radars
2. Ingest stream gauge data (real-time and forecast)
3. Ingest inundation maps
4. Ingest network structure of river system and communities
5. Ingest spatial layers for watershed boundaries, river network, and various map layers
6. Create dynamic visualizations of stream level data
7. Create interactive visualizations of flood inundation maps
8. Create rainfall intensity animation and accumulation
9. Create community view using the network structure and connectivity scheme
10. Update all data and visualization every 5 to 20 minutes

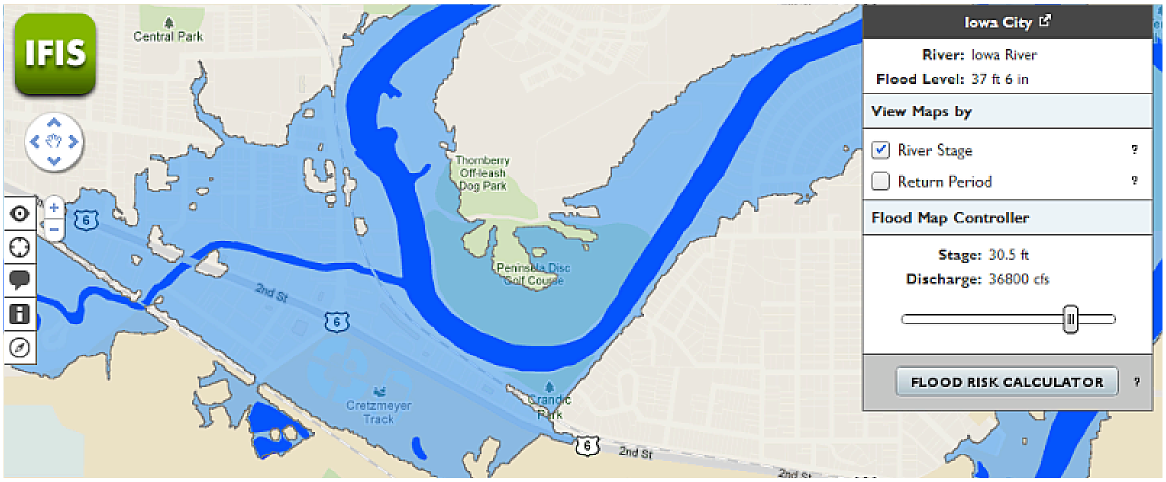


IOWA CITY (IOWA RIVER)

Population 67831
Land Area 22 sq mi

WATERSHED CHARACTERISTICS

Upstream Area 3268 sq mi ?
Travel Time 7 day ?
Rainfall Coverage trace ?
Max. Rainfall 0 in/h ?



Iowa City

River: Iowa River
Flood Level: 37 ft 6 in

View Maps by

☑ River Stage ?
☐ Return Period ?

Flood Map Controller

Stage: 30.5 ft
Discharge: 36800 cfs

FLOOD RISK CALCULATOR ?

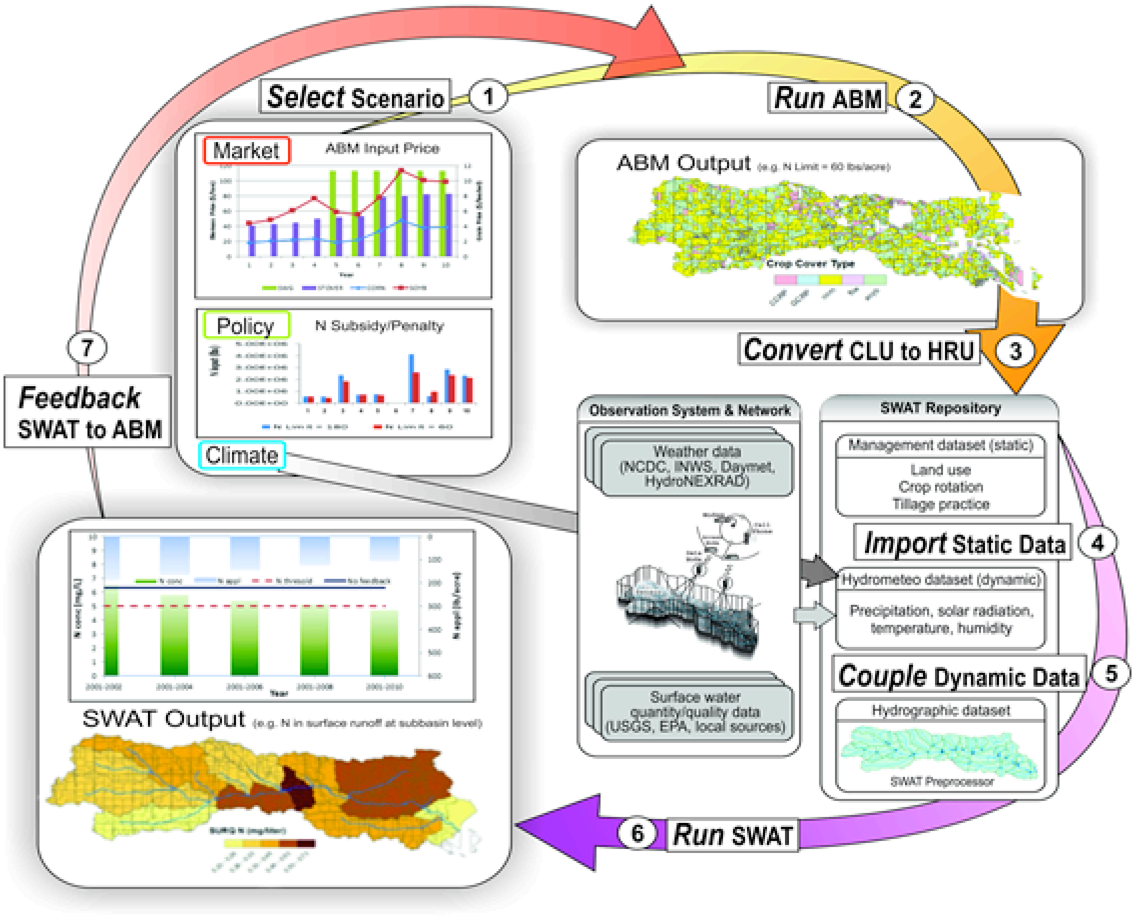
This project is developed at the Iowa Flood Center (IFC). Science POC: Witold F. Krajewski UIOWA). Technical POC: Ibrahim Demir (UIOWA)

A Whiteboard Workflow Paradigm Highlight:
Intelligent Digital Watershed (IDW)

IDW enables users not only to access, store, and display a variety of heterogeneous data, but also to understand the links between shifts in crop choices and soil conservation practices and the water quantity and quality in watershed streams as well as the public perception of environmental health. For this purpose an individual-level Agent-based Models (ABM) simulates land-use decision-making processes under alternative scenarios about: a) agricultural and environmental policies; b) market forces associated with biofuel production. Modeled decisions are linked to a watershed simulation (Soil & Water Assessment Tool - SWAT) model to understand the impact these scenarios may have on indicators of water quality (nitrate, phosphate, dissolved oxygen). The cyber-enabled platform enables better understanding between the expanding biofuel economy, land management, and water quality. The workflow can support researchers in transforming data into knowledge about interrelated socio-economic and biophysical process as well as stakeholders for making informed decision making. The current system is not utilizing a standardized workflow technology. The workflow can be accessed online at http://iowadis.org/idwdev.

Workflow Steps

1. Select scenario (by choosing crop prices, subsidies for fertilizers, and climate type)
2. Run ABM (to provide crop choice, fertilizer & pesticide application rates and application time, type of soil conservation practice, etc)
3. Convert Common Land Unit (farm boundaries) to Hydrologic Response Unit (generated by the SWAT model)
4. Import static data (DEM, soil type, ABM output)
5. Couple Dynamic Data (ingest time series for precipitation and temperature – raw data records or as altered in Step 1)
6. Run SWAT (to provide water quantity and quality at selected watershed points)
7. Feedback SWAT to ABM (concentrations of pollutants and suspended sediment)


Select Scenario
1
Run ABM
2
Market
ABM Input Price
Policy
N Subsidy/Penalty
Climate
ABM Output (e.g. N Limit = 60 lbs/acre)
Crop Cover Type
Convert CLU to HRU
3
Observation System & Network
Weather data (NCDC, INWS, Daymet, HydroNEXRAD)
Surface water quantity/quality data (USGS, EPA, local sources)
SWAT Repository
Management dataset (static)
Land use
Crop rotation
Tillage practice
Import Static Data
4
Hydrometeo dataset (dynamic)
Precipitation, solar radiation, temperature, humidity
Couple Dynamic Data
5
Hydrographic dataset
SWAT Preprocessor
Run SWAT
6
SWAT Output (e.g. N in surface runoff at subbasin level)
SURQ N (mg/liter)
Feedback SWAT to ABM
7

This workflow is a collaborative effort integrating scientists from hydro- and geoinformatics, hydrology, computer science, civil & environmental engineering at University of Iowa (UI) with environmental & energy economics investigators from Southern Illinois University. Science POC: Dave Bennett (UI). Technical POC: Marian Muste (UI).

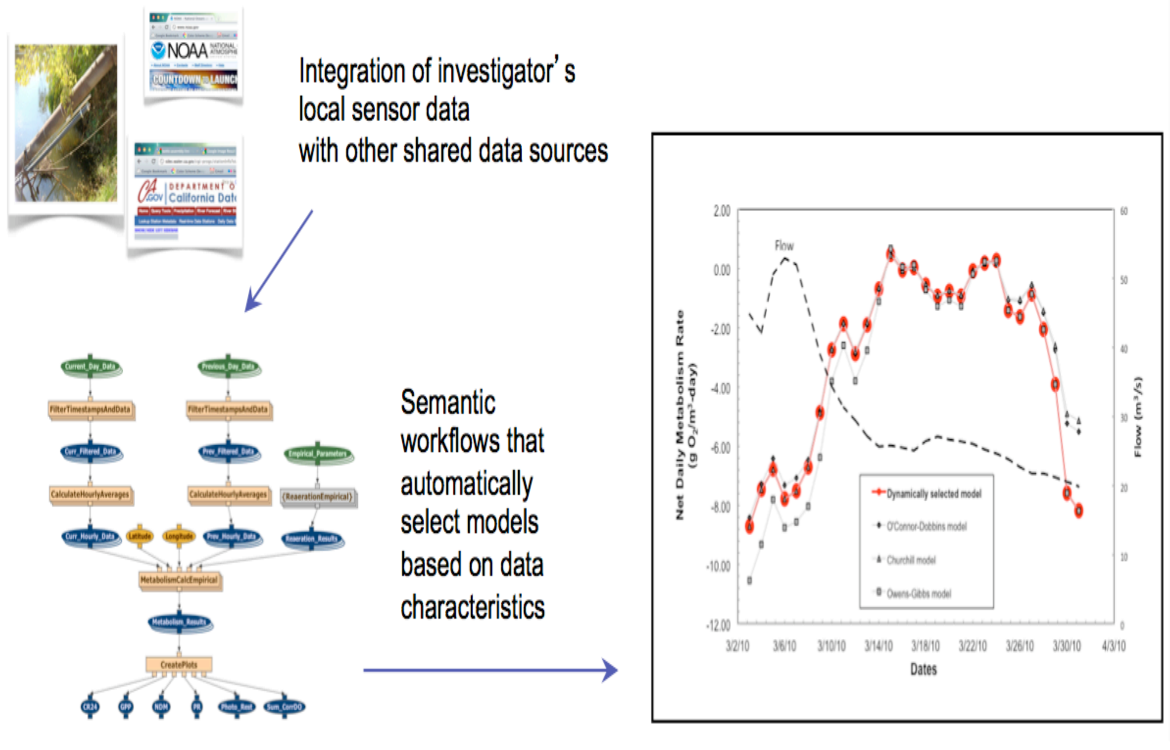
A Long Tail Workflow Paradigm Highlight:
Efficient Data Analysis through Automated Model Selection

Adaptive modeling frameworks for data analysis can be created with workflow systems. Semantic metadata and provenance are used throughout the data analysis process in the Karma data integration and the Wings workflow system to automatically choose models for water reaeration depending on river flow conditions. Karma integrates data from sensors with data from regional and national sources, generating metadata that is attached to the integrated datasets. Wings uses the flow, velocity, and reach geomorphology for each day of the period of analysis to choose an appropriate model, effectively configuring a different workflow to be run every day. More details about this work can be found in a published article, and about the Wings workflow system at http://wings-workflows.org.

The benefit of this system to the scientist is the dynamic selection of models depending on the characteristics of the data. The graphic below shows on the right hand side the daily metabolism rates generated by four different reaeration correction models. The red dots highlight the model selected dynamically by the system, which changes over time and is optimal for different flow conditions. This is achieved through the combination of two provenance-aware systems, one for data integration (Karma) and another for data analysis (Wings). Karma generates semantic metadata during data integration, which is used by Wings to select the appropriate model for the executed workflow.

Workflow steps:

1. Data preparation and integration
    a. Retrieve flow data from sensors available at the California Data Exchange Center (CDAC) site (water.data.gov) (flow, river discharge, electrical conductivity, etc).
    b. Integrate with local sensor data about water quality (dissolved oxygen, temperature, specific conductivity, depth, etc).
    c. Integrate meteorology data from NOAA site.
2. Data analysis
    a. Calculate hourly averages
    b. Obtain reaeration estimates
    c. Estimate community respiration
    d. Estimate gross primary productivity
    e. Calculate net daily metabolism
3. Create plots for relevant results of the analysis



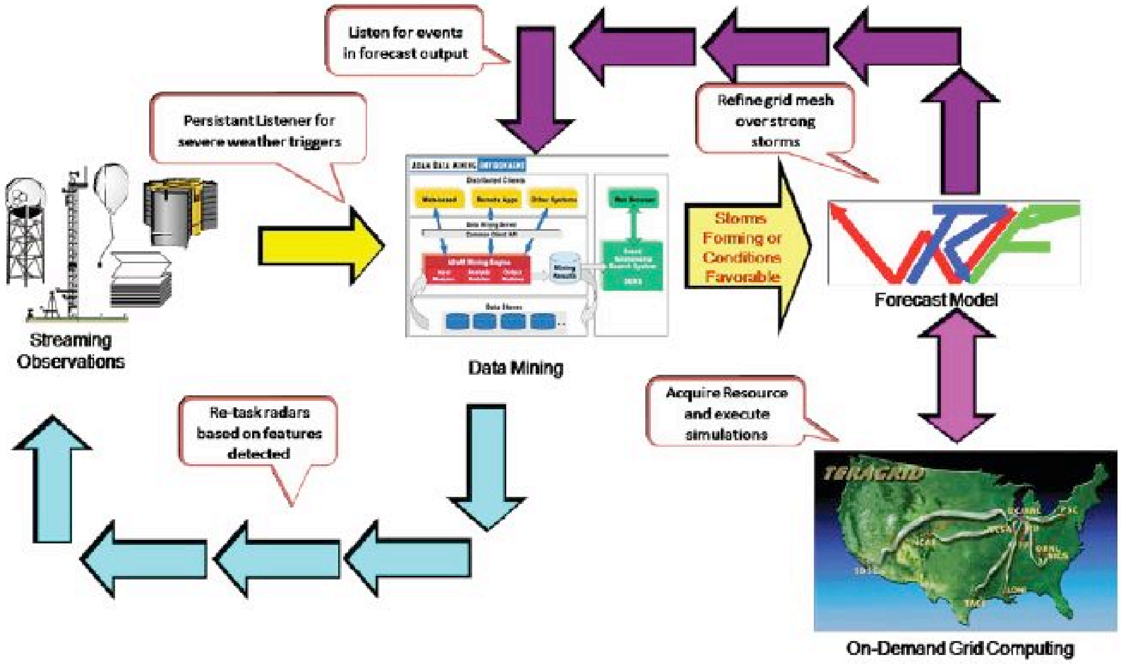
Integration of investigator's local sensor data with other shared data sources

Semantic workflows that automatically select models based on data characteristics

This research is a collaboration between computer scientists at USC and hydrologists at UC Merced. Science POC: Tom Harmon (UC Merced). Technical POCs: Yolanda Gil (USC), Craig Knoblock (USC), and Pedro Szekely (USC).

A Big Data Paradigm Highlight:
The Grand Challenge Linked Environment for Atmospheric Discovery Workflow

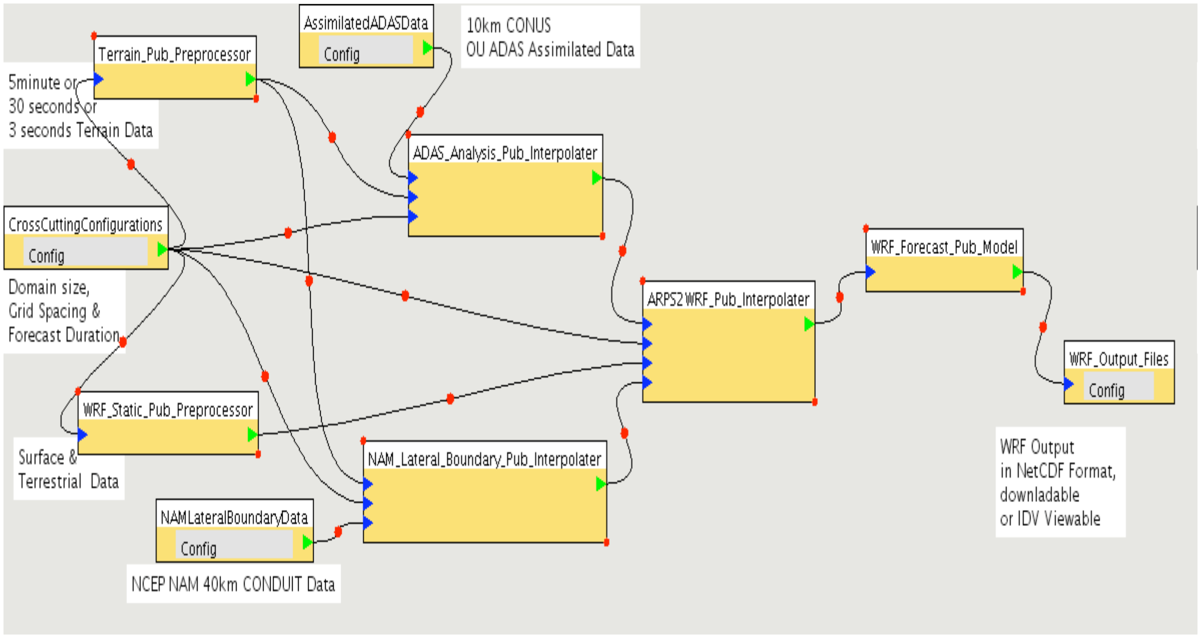
This system was envisioned to investigate how we could better understand mesoscale atmospheric events such as severe storms through real-time data stream mining and better-than-real-time weather forecasting by adapting our technologies and approaches to the weather as it occurs. The dynamic workflow of this system is illustrated below.


Listen for events in forecast output
Persistant Listener for severe weather triggers
Refine grid mesh over strong storms
Streaming Observations
Data Mining
Storms Forming or Conditions Favorable
Forecast Model
Acquire Resource and execute simulations
Re-task radars based on features detected
TERAGRID
On-Demand Grid Computing

Workflow Steps

1. Collection of streaming data from various meteorological sources.
2. Continuously run sophisticated data mining algorithms to extract potential mesoscale signatures.
3. Trigger weather forecasting workflows in areas of detected storms.
4. Acquire, Assimilate, Interpolate observation data onto target model grids and prepare initial and lateral boundary conditions.
5. Launch compute intensive ensemble models on available National Computational cyberinfrastructure.
6. Generate post processed outputs and visualizations.
7. Dynamically refine the model configurations, ingested data based by mining the initial coarser model execution.
8. Optionally interact with adaptive sensors in fetching more localized focused observations.
9. Repeat the steps, predicting the weather as it occurs.

This system leverages *Apache Airavata* (http://www.airavata.org/), which is an open source software system to build science gateways. The diagram below depicts the workflow steps codified in the Apache Airavata. workflow management tool.



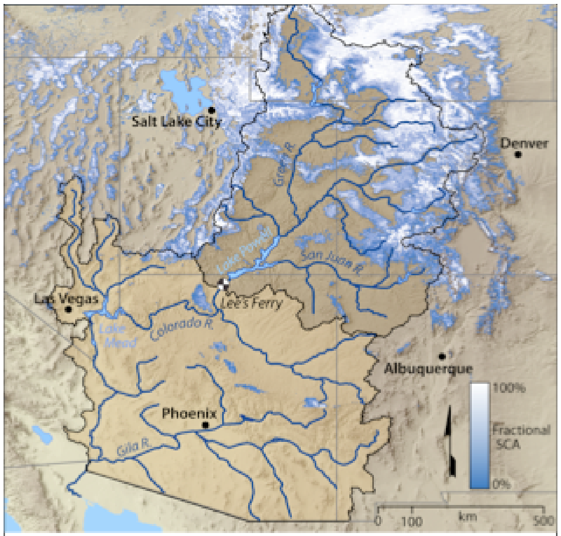
This workflow vision is a collaborative effort by a multidisciplinary team from OU, IU, NCSA, Unidata, UAH, Howard, Millersville, Colorado State, and RENCI, and included significant education and outreach.

A Center Modeling Paradigm Highlight:
Scientific Data Management and Processing System for Seasonal Forecasts of Snowmelt

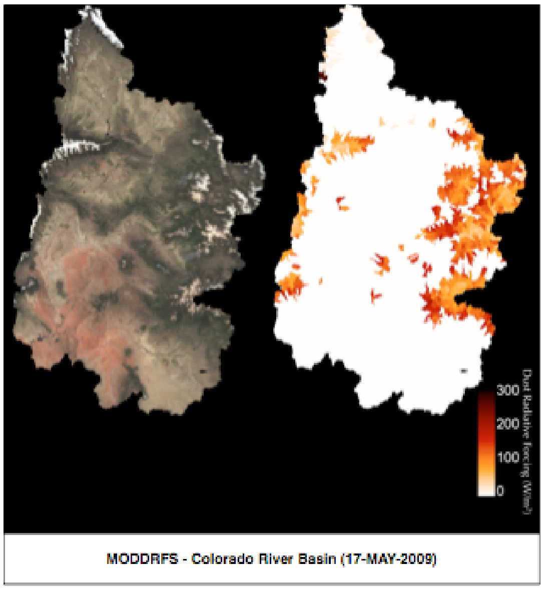
Snow cover and its melt dominate regional climate and hydrology in many of the world's mountainous regions. One-sixth of Earth's population depends on snow or glacier-melt for water resources. Operationally, seasonal forecasts of snowmelt-generated streamflow are leveraged through empirical relations based on past snowmelt periods. These historical data show that climate is changing, but the changes reduce the reliability of the empirical relations. Therefore, optimal future management of snowmelt derived water resources will require explicit physical models driven by remotely sensed snow property data. Toward this goal, the Snow Optics Laboratory at the Jet Propulsion Laboratory (JPL) has initiated a near real-time processing pipeline to generate and publish post-processed snow data products within a few hours of satellite acquisition. To solve this challenge, a Scientific Data Management and Processing System was required and the JPL Team leveraged an open-source project called Apache Object Oriented Data Technology (OODT).

Workflow Steps

1. Retrieve MODIS data from Land Processes Distributed Active Archive (LPDAAC) and place in staging area.
2. Crawl staging area to extract and catalog metadata and ingest data into archive.
3. Pull out bands from MODIS data that will be used from processing.
4. Run MODSCAG to produce Snow Covered Area and Grain Size products and ancillary data.
5. Produce GeoTIFFs from raw binary data.
6. Extract and catalog metadata from MODSCAG outputs and ingest data into archive
7. Run MODDRFS to produce Dust radiative forcing on snow.
8. Produce GeoTIFFs from raw binary data.
9. Extract and catalog metadata from MODDRFS outputs and ingest data into archive.
10. Reproject outputs and register with GIS server.
11. Repeat the steps above on a daily basis.



MODSCAG - Colorado River Basin



MODDRFS - Colorado River Basin (17-MAY-2009)

This work is a collaborative effort between snow hydrologists and computer scientists at JPL, University of Utah, and NSIDC. Technical POC: Chris Mattmann (JPL),. Science POC: Thomas Painter (JPL).

A Long Tail Workflow Paradigm Highlight:

Linking Scientific Models to People: Decision Support Engine for Participatory Modelling Experiments and Engagement

Purpose / Application: The Groundwater Decision Support System (GWDSS) was conceived as an interactive system to determine sustainable yield for aquifer systems. The larger GWDSS included modules for numerical groundwater modelling, systems dynamics, and a non-classical optimization algorithm. GWDSS has supported participatory modelling exercises for various case studies and also supports research into uncertainty for groundwater. The initial, or alpha, case study was completed for a rapidly urbanizing area in central Texas United States. The groundwater system is a karst sytem that is senstive to land use change. The integrated models were developed in a participatory process with stakeholders and groundwater managers to evaluate alternative scenarios for managing allocation and drought policies.

End-users: Earth resource planning agencies and community stakeholders are the primary end users of the GWDSS. However, it can still be used by scientists to evaluate uncertainty in policy and management recommendations. GWDSS has also become useful for STEM Education efforts and teaching integrated modelling skills.

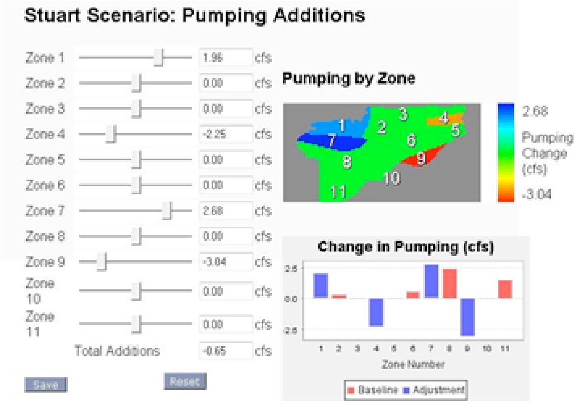
Brief Model Description: The original case inside GWDSS is built around a science-vetted numerical groundwater model. The system dynamics components are linked to the groundwater model as elements change the inputs into the groundwater system (e.g. land use modified recharge) and modelled outputs from the groundwater system as feedbacks to the system dynamics components (e.g. drought conditions trigger pumping cutback policies which then change pumping in the model).

Implementation: GWDSS is implemented as Tomcat/Apache/MySQL stack using the primary programming language. The Java “middleware” connects groundwater models (MODFLOW written in fortran), Powersim system dynamics models using the proprietary SDK provided by the vendor, and a custom optimization and search algorithm (tabu) that was implemented as .jar file for execution. There are both desktop and web-enabled versions of GWDSS. The GWDSS systems is used for the group dialogue and participatory sessions with groups because system dynamics can run simulations rapidly, and support “what-if’ discussions in real-time. Also, system dynamics allowed integrated aspects (e.g. land use change, economics, demographics) to be represented in the system.

Workflow Steps

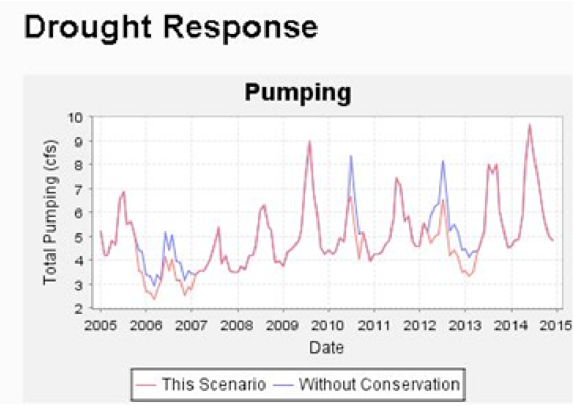
1. Develop a set of problem frames through elicitive stakeholder interviews
2. Complete narrative and value focused thinking analyses to define objective functions, constraints, decision variables, and links to numerical simulation models (e.g. in the case of hydrogeology these may include water levels, flows, saturated thickness, etc)
3. Set up basic case information in database and file-based directories
4. Connect groundwater models and systems dynamics models of ancillary elements into mysql database
5. Formalize decision and objective problem formulation
6. Assure that all parameters needed to match with problem formulations are ‘exposed’ and accessible
7. Link problem formulations and parameters/variables with simulation-optimization engine
8. Begin scenario testing and search for satisficing solutions

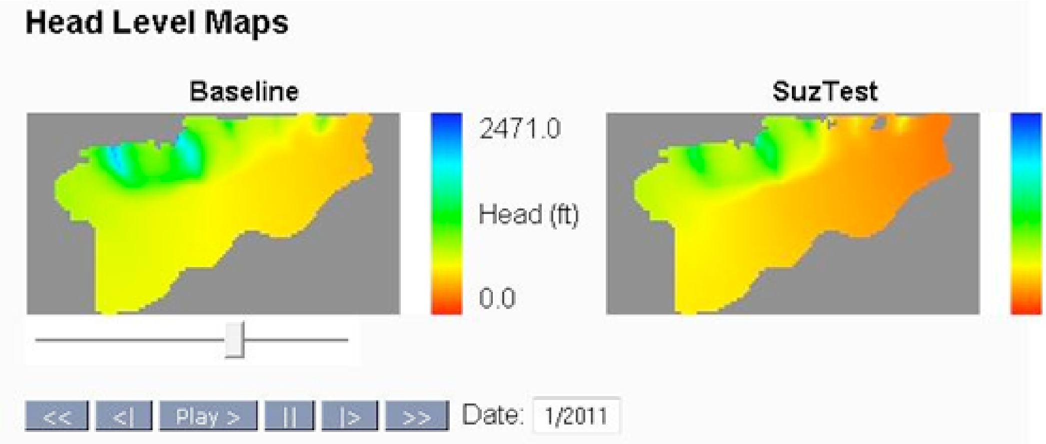
9. Design/implement case specific visualizations in preparation for stakeholder working sessions
10. Conduct interactive dialogue sessions with participating stakeholders



Stuart Scenario: Pumping Additions

Zone 1	1.96	cfs
Zone 2	0.00	cfs
Zone 3	0.00	cfs
Zone 4	-2.25	cfs
Zone 5	0.00	cfs
Zone 6	0.00	cfs
Zone 7	2.68	cfs
Zone 8	0.00	cfs
Zone 9	-3.04	cfs
Zone 10	0.00	cfs
Zone 11	0.00	cfs
Total Additions	-0.65	cfs





This work is a collaborative effort of Suzanne A. Pierce (PI) The University of Texas at Austin, Anthony Jakeman Australian National University, Sondoss El Sawah, Australian National University, Eugenio Figueroa Universidad de Chile, Craig Simmons Adelaide University, and Roy Jenevein, The University of Texas at Austin.

A Long Tail Workflow Paradigm Highlight:
ACE: Age Calculation Engine

ACE is a software design environment built to calculate landform ages using cosmogenic nuclide dating. It includes calibration and dating algorithms for 3He, 10Be, 14C, 21Ne, 26Al and 36Cl. ACE makes use of workflow technology to allow these calibration and dating algorithms to be extended easily to additional nuclides and performed on a broad range of data sets.

Before ACE, researchers in this community made use of spreadsheets or other specialized tools that had assumptions baked in and were practically impossible to customize. Workflow was used as a means to move the geoscientists away from these hardwired spreadsheets to small modules of Python-based code that could then be wired together using a workflow specification and executed on various data sets. We arrived at workflow technology after an initial attempt to make use of a Java-based program that implemented ACE’s dating algorithm. The style imposed by Java’s syntax and its numerical classes was rejected by our geoscience colleagues as being too unfamiliar. We then introduced them to Python, showed them how the syntax was very similar to what they were used to in Excel, and helped them convert their Excel spreadsheets into small Python functions that could then be linked together using workflow technology.

This approach took advantage of the considerable computational skills that geoscientists possess while incrementally moving them to a state in which their calculations could be customized by swapping one python module for another via a simple change in our workflow editor. To demonstrate the coding skills of our geoscientist colleagues, consider the types of equations that appeared in their spreadsheets:

```
=IF (B102<4000,
       VLOOKUP(FLOOR(B102,500),paleomag_table,3) +
    (B102-(FLOOR(B102,500))) *
    (((VLOOKUP(CEILING(B102,500),paleomag_table,3)) -
       (VLOOKUP(FLOOR(B102,500),paleomag_table,3))) /
       (CEILING(B102,500)-(FLOOR(B102,500)))),
       VLOOKUP(FLOOR(B102,1000),paleomag_table,3) +
    (B102-(FLOOR(B102,1000))) *
       (((VLOOKUP(CEILING(B102,1000),paleomag_table,3)) -
       (VLOOKUP(FLOOR(B102,1000),paleomag_table,3)))/
       (CEILING(B102,1000)-(FLOOR(B102,1000)))))
```

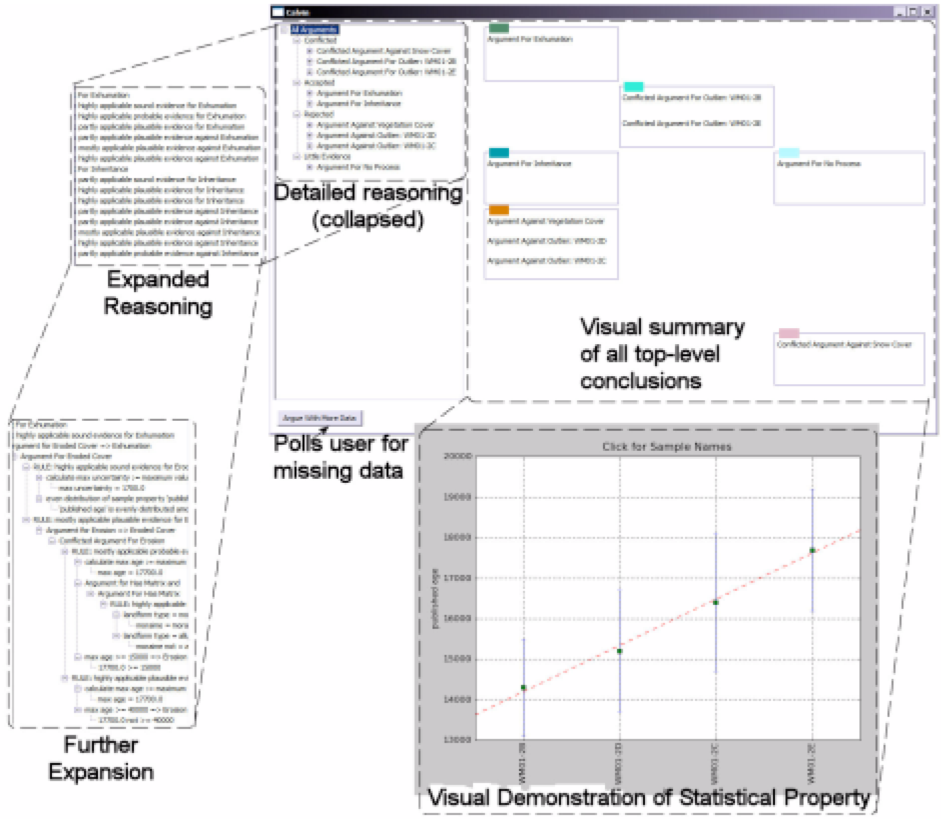
This line uses conditional constructs, lookup tables, and non-trivial mathematical constructs to determine the age of a rock sample, based on the presence of certain cosmogenic nuclides. The spreadsheet that contained this code consisted of many different pages that contained elementary data sets, complex sub-calculations that were accessed via lookup tables from the primary calculations, and circular references that forced the spreadsheet to iteratively process the computation until the change in a particular value converged to a specified threshold.

Over the course of the ACE project, the geoscientists became comfortable with the incremental and modular nature of workflow technology and began to write code that was smaller and easier to understand and maintain:

```
def calculateQ_s(self, s):
        # calculate (LAMBDA_f/Zs)*(1-EXP(-1*Zs/LAMBDA_f))
        div1 = self.LAMBDA_f / s["Zs"]
        div2 = -s["Zs"] / self.LAMBDA_f
        sub1 = 1.0 - math.exp(div2)
        s["Q_s"] = div1 * sub1
```

Each of these mini-calculations were wrapped in a workflow component and then arranged into algorithms via a custom workflow editor. This editor allowed the scientists to specify the ordering of these small python components such that they would perform a calibration or calculate ages on a specified dataset. Our engine supported workflows that contained both branches and loops; it was implemented to continue executing the workflow until all of the rock samples in the data set had been processed.

The output of these calculations could then be displayed in a *sample viewer* or be fed into a forensic reasoning system that would generate the output shown in the figure below to help the geoscientists understand the history of the landform.


All Arguments
Conflicted
Conflicted Argument Against Snow Cover
Conflicted Argument For Outlier: WM01-2B
Conflicted Argument For Outlier: WM01-2C
Accepted
Argument For Exhumation
Argument For Inheritance
Rejected
Argument Against Vegetation Cover
Argument Against Outlier: WM01-2D
Argument Against Outlier: WM01-2C
Little Evidence
Argument For No Process
Argument For Exhumation
Conflicted Argument For Outlier: WM01-2B
Argument For Inheritance
Argument For No Process
Argument Against Vegetation Cover
Argument Against Outlier: WM01-2D
Argument Against Outlier: WM01-2C
Conflicted Argument Against Snow Cover
Argue With More Data
Detailed reasoning (collapsed)
Expanded Reasoning
Further Expansion
Visual summary of all top-level conclusions
Polls user for missing data
Click for Sample Names
published age
20000
19000
18000
17000
16000
15000
14000
13000
WM01-2B
WM01-2D
WM01-2C
WM01-2E
Visual Demonstration of Statistical Property

ACE was jointly developed by researchers at the University of Colorado at Boulder and the University of Arizona as part of the work performed on NSF grants ATM-0325812 and ATM-0325929. The software is available at <http://ace.hwr.arizona.edu/>. For questions concerning the science behind cosmogenic nuclide dating, please contact Marek Zreda

<marek@hwr.arizona.edu>. For questions concerning ACE’s workflow engine and its software capabilities, please contact Elizabeth Bradley <elizabeth.bradley@colorado.edu> and Kenneth M. Anderson <kena@cs.colorado.edu>.

A Long Tail Workflow Paradigm Highlight:
Real-Time Science Mission Situational Awareness Workflow

The NCAR Mission Coordinator System (MCS) integrates real-time position information and science observations from multiple aircraft, supplementary observations, other agency real-time data products (e.g. NWS NEXRAD), weather forecasts, and mission plans. The system provides a comprehensive and comprehensible overview of the current state of an operational atmospheric science field campaign. The critical users are onboard aircraft scientists and ground based mission managers. Project scientists, instrument engineers and support technicians are additional stakeholders. Technology areas include real-time communications, satellite data links, data scraping, format conversion, digital product routing, database manipulation, web services, and visualization.

The MCS ultimately satisfies two needs: comprehensive up-to-date situational awareness, and real-time project communications. The system was built upon an ad hoc suite of software approaches, such as shell scripts, cron job schedulers, databases, web apps, and standalone applications.

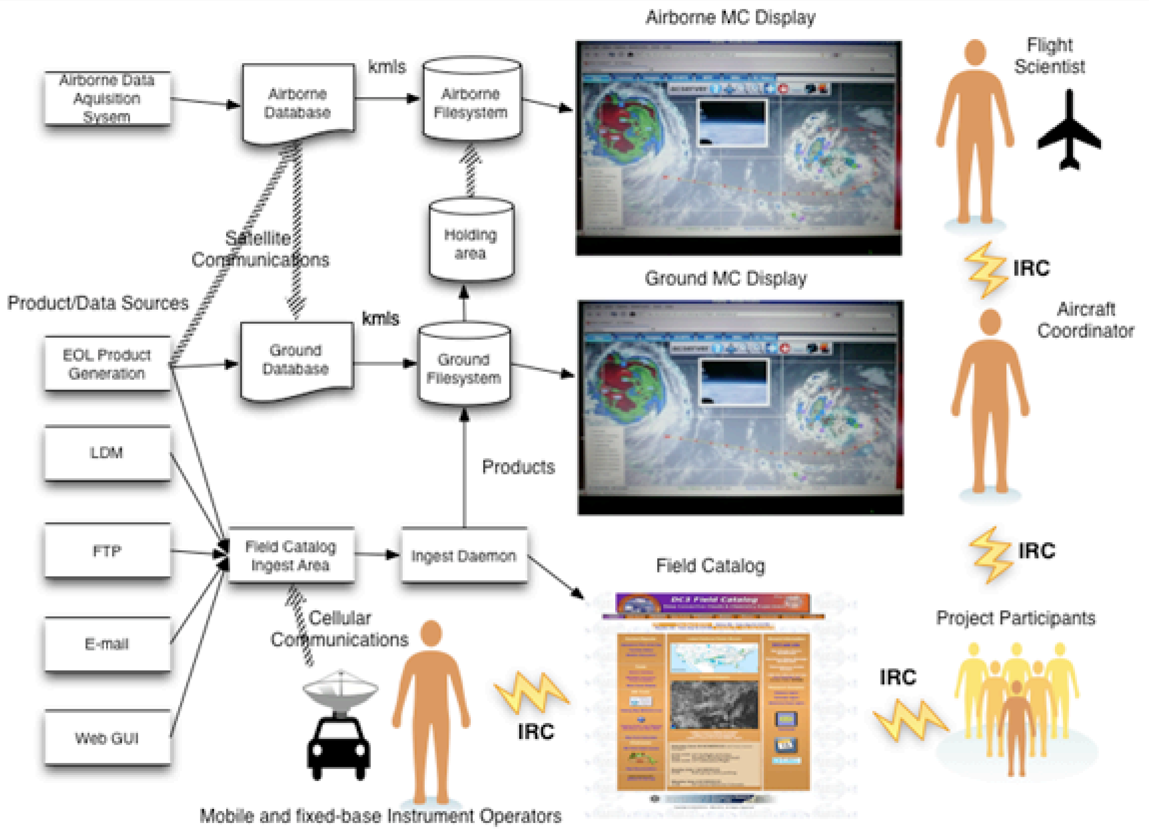
One of the greatest challenges has been reconfiguration and adaption for each successive deployment, as the product types and sources vary greatly between each science mission. For example, one project might be investigating the role of aerosols in the chemistry of the marine boundary layer, and the following project would involve aircraft observations in close proximity vigorously developing thunderstorms. Each calls for a different collection of support products and visualizations. The number and capability of aircraft change between projects, as do the scale and functions of ground based mission support facilities.

The MCS does not currently employ a workflow system. A workflow architecture could be extremely useful in organizing and easing system reconfiguration, data communications, and data management. System robustness would be enhanced as well.

Workflow Steps

1. Characterize, enumerate and automate collection and/or generation of source data product types. (E.g. NEXRAD observations, lightning network maps, satellite imagery, model forecasts, etc.)
2. Collect real-time aircraft position and track information.
3. Collect real-time mobile research platform location information.
4. Push important products for real-time decision-making to the aircraft and field catalog.

5. Decision makers on the ground and in the air monitor the MCS for changing conditions and nowcast information and communicate via IRC chat to adjust flight plans based on best sampling options for targets of opportunity and hazard avoidance.
6. Ground-based fixed and mobile instrument operators monitor changing conditions and aircraft plans via the field catalog tools and IRC chat and adjust sampling strategies to coordinate sampling per project decisions.
7. Project participants and others interested in the campaign monitor field operations via the field catalog and its associated tools.



Point of Contact: Mike Daniels, <daniels@ucar.edu>, Computing Data and Software Facility, NCAR Earth Observing Laboratory; Charlie Martin, <martinc@ucar.edu>, Computing Data and Software Facility, NCAR Earth Observing Laboratory; Greg Stossmeister, <gstoss@ucar.edu>, Computing Data and Software Facility, NCAR Earth Observing Laboratory

A Long Tail Workflow Paradigm Highlight:
NCAR/Earth Observing Laboratory Field Project Data Services Lifecycle Workflow

The following Steps describe the workflow that NCAR Earth Observing Laboratory (EOL) performs to support a variety of data services for its field projects. These Steps follow the lifecycle of a field project from initial planning to the final archive.

Workflow Steps

1. Involvement in the early project data planning
2. Data Questionnaire
3. Collection of PI and supporting datasets
4. Supporting Project Participants in the field
5. Develop and provide an on-line Field Catalog
6. Data quality control and processing

1) Involvement in the early project data planning – EOL provides early advice and coordination with project Principal Investigators (PIs) to develop a data management strategy, data policy (in some cases international), and exchange protocol coordinating with project partners. In many cases this involves collaboration with other data archives and data providers. EOL also develops and maintains project web pages that contain all project information including project description, deployment logistics, documentation, meetings, presentations, mailing lists, contacts, and related links. This helps with the coordination of pre-field deployment issues and sharing of information. EOL then prepares a Data Management Plan document that contains all these details.

2) *Data Questionnaire* –If necessary, EOL develops and distributes an on-line data questionnaire to query the project PIs on what datasets and products they will be submitting to an on-line Field Catalog (see Step 5 below) and final archive as well as what supporting data will be needed to be archived for the post-field analysis. The results of the questionnaire (or discussions with the project participants) are summarized and used to define the Field Catalog and final archive structure and content. Additional questionnaires are prepared as needed to support field planning.

3) *Collection of PI and supporting datasets* – EOL develops the appropriate agreements, arrangements, and data ingest mechanisms to populate the data, products, and metadata for the Field Catalog and long-term archive. Depending on the data sources, some of these datasets have to be archived later when available. This is particularly true for the PI research datasets where post-field processing and further quality assurance are required following the field deployment. EOL develops and supports data acquisition systems for NSF's Lower Atmospheric Observing Facilities (LAOF) such as radars, aircraft, profilers, surface and sounding systems. For some platforms, EOL's data acquisition systems also serve as a central hub that connects to University-developed instrumentation onboard aircraft, for example. EOL also provides near real-time support for ingest of research upper air soundings into the Global Telecommunications System (GTS) for ingest by international numerical Weather Prediction Centers (e.g. ECMWF, NCEP, JMA, etc.). This effort involves ingesting high-resolution data from each field source, re-formatting the data, preparing the GTS message, and submitting these messages through the U.S. NWS Telecommunications Gateway.

4) *Supporting Project Participants in the field* - In preparation for the field deployment, EOL develops and implements a data communications network for the Operations and Analysis Center(s) and performs testing to determine available bandwidth and connection reliability between Boulder and the Field Facilities. EOL provides on-site support for this network during the campaign and assists individual users with internet connectivity using cellular communications technology as well.

5) *Develop and provide an on-line Field Catalog* - The EOL Field Catalog is a web-based tool that allows the project participants to post (and access) operations and mission/scientific reports, operational and preliminary research imagery/products (e.g. satellite, surface, upper air, radar), model output fields, and project documentation. The Field Catalog is used to make operational decisions during the field phase as well as provide a project summary and “browse” tool for use by researchers in the post-field analysis phase. EOL customizes the Field Catalog for specific project requirements and begins operating it approximately 1 month prior to field deployment. This allows enough time for the PIs to review the products and make any modifications as needed before field operations begin. EOL monitors and maintains the Field Catalog through the duration of the deployment and also provides in-field support and training to the project participants. The Field Catalog remains as part of the long-term project archive.

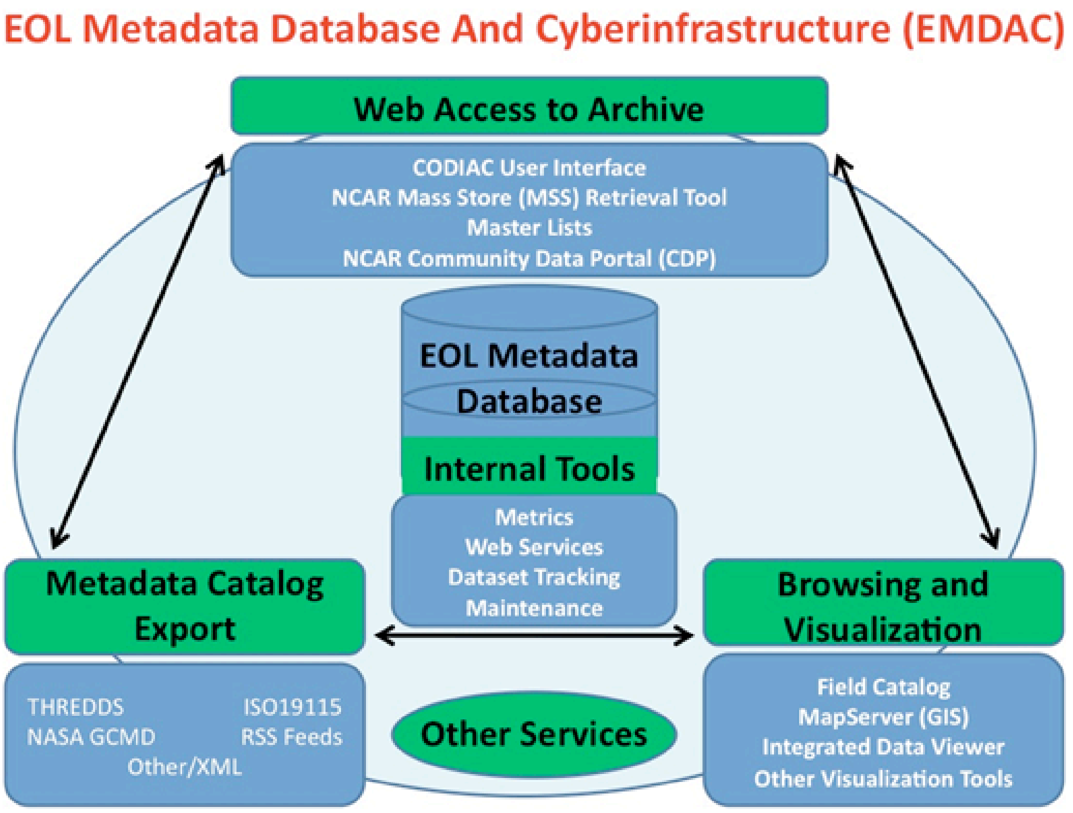
EOL has recently developed a multi-way mirrored catalog system that allows products and reports generated for the project at remote locations to be sent to a Field Catalog at NCAR (Boulder) and then relayed based on a priority system to other local catalogs with varying levels of bandwidth (e.g. ships). It also allowed users to quickly upload products and reports to the Field Catalog without having to deal with delays in uploading information remotely around the world.

Another catalog feature includes a “web forum”. As participants in a field deployment are distributed across so many different time zones and the need for instant communication to make real-time decisions are limited, people can post and read messages at their leisure on a variety of scientific and operational (e.g. forecasting, logistics, instrumentation, travel, etc.) topics of interest. The forum then becomes part of the archive following the field deployment.

Another EOL tool, the Mission Coordination System (MCS), provides GIS overlay capability to produce integrated, real-time products as part of the Field Catalog. This tool is useful in overlaying such layers as aircraft flight tracks, dropsonde locations, sounding plots, or other operational and research data such as radar, satellite, and lightning strike data. The MCS provides not just an ability to overlay and view these products in real-time but also the capability to playback these products from any particular date and time when data was collected.

6) *Data quality control and processing* – EOL works with project PIs to determine requirements for special or value-added dataset generation. EOL performs Quality Control (QC) on data from its own LAOF platforms and also collects all supporting, highest resolution, surface and sounding data in the field project domain (e.g. National and Regional meso-networks). EOL performs various levels of automated and manual QC on these data and including time series analysis, horizontal and vertical consistency checks, comparison with other data sources, and visual examination. As datasets are ingested at EOL from a variety of sources (e.g. LDM, FTP, e-mail, media, etc.), they are logged to the Data Tracking System (DTS). The DTS is a collection of tools (i.e. data loading notes, processing inventory tool, and project statistics/metrics) used to track data and metadata status and versioning from ingest to final archive. From these data, EOL can also create both surface and upper air sounding "composite" data sets. The surface composite data set involves the collection of all operational/research surface network data from all available sources, extraction of common standard meteorological parameters, conversion of all data to a common format, provision of uniform QC, and generation of final composite data sets at various time resolutions (e.g., 5-min and hourly). Similar to the process outlined above for surface data, upper air composites can be generated from operational and research rawinsonde and dropsonde data. These composites generally include: (1) All soundings in highest vertical resolution; and (2) All soundings interpolated to 5-mb levels (important for use in model ingest, initialization, and comparison). The major advantage of creating these surface and upper air composite data sets is cost efficiency by eliminating the requirement of each investigator to individually re-process and re-format separate network data sets. In addition, all native resolution and formatted data are also made available as part of the project archive prior to the completion of the composite data sets.

7) Establish a final archive - It is important to the project researchers and the scientific community to establish a final long-term project archive that is easy to access. This archive generally contains all PI collected data/metadata as well as supporting datasets (e.g. satellite, surface, upper air, model output) required for the post-field analysis phase. Project data are primarily stored on the NCAR High Performance Storage System (HPSS) and accessed through the EOL Data Management System (EMDAC) as shown in Fig. 1.



EMDAC provides comprehensive support for investigator data submission and acquisition, data access control, metadata generation, project dataset inventory and tracking, metadata catalogs, dataset archival and long-term stewardship. EOL provides links to multi-agency data portals, and creates compatible metadata and data access infrastructure through a modular and extensible architecture. The EMDAC has the capability that allows the export of metadata in a variety of formats (e.g. ISO19115, THREDDS, XML, etc.). This allows for project datasets to be discoverable through a variety of other catalogs and portals such as the NASA Global Change Master Directory (GCMD).

EOL develops and implements a series of project data management web pages to provide access to all distributed project data sets, documentation, on-line Field Catalog products, collaborating project data archives, data submission guidelines and instructions, and other relevant data links for the long-term archive. EOL provides “one-stop shopping” for all datasets through a web-based “Master Project Dataset List” that provides links to access all project data/metadata at EOL and other distributed related project datasets. This Master List is continually updated as new datasets become available. Password protection can be applied to those datasets in accordance with the project data policy and data provider restrictions. In addition, EOL (in coordination and consultation with the project PIs) develops and maintains a publications list as the project enters the post-field data analysis phase.

EOL also provides support to hosting or organizing project scientific meetings and data workshops. In many cases this involves providing specialized data services such as data servers allowing the participants to access and work with data during the Workshop to foster collaboration.

Point of Contact: Mike Daniels, <daniels@ucar.edu>, Computing Data and Software Facility, NCAR Earth Observing Laboratory; Steve Williams, <sfw@ucar.edu>, Computing Data and Software Facility, NCAR Earth Observing Laboratory; Greg Stossmeister, <gstoss@ucar.edu>, Computing Data and Software Facility, NCAR Earth Observing Laboratory

1.4 Technical Areas Addressed by this Roadmap

Scientific workflows are a central element of modern cyberinfrastructure, typically used to describe, compose, model, and execute ensembles of scientific computations, data analyses or visualizations on distributed resources (Deelman and Gil ed., 2006, Gil, 2009). The prescriptive and descriptive representations of workflows (called provenance) will be useful for publishing, discovering, and reproducing computational results in geosciences. Scientific workflows within the geosciences range from pipelines used to create community data products to real-time processing of sensor data, to individual researchers' unique computations. Current status of the state-of-the-art in geoscience is discussed in detail in Section 5, "Status".

The development and deployment of workflow technologies encompasses a broad range of challenging topics including distributed execution management, the coupling of multiple models into composite applications, the integration of a wide range of data sources (ranging from real-time to archival data) with processing, and the creation of refined data products from raw data. Additional challenges concern the interaction of scientists with workflow systems, so that workflows can be easily created, validated, published, visualized, and reused. See Sections 3, "Challenges" and Section 4, "Requirements" of this roadmap for more information.

Many requirements in this roadmap require fundamental research in computer science. Part of the work to be done under this roadmap is to identify these basic research areas in coordination with computer science researchers, and to establish priorities for EarthCube. Computer science researchers need to be brought into the EarthCube community to participate as they will be crucial contributors to the accomplishment of the goals in this roadmap.

1.5 Improvements to the State-of-the-Art Enabled by this Roadmap

The overarching goal of the workflow roadmap is to make workflow technologies and software ubiquitous within the geosciences. Workflow researchers have already investigated a wide range of problems relevant to geoscientists. We must now work together to improve mutual understanding of requirements and challenges, increase adoption, and develop governance models that will broaden commitment to community software solutions.

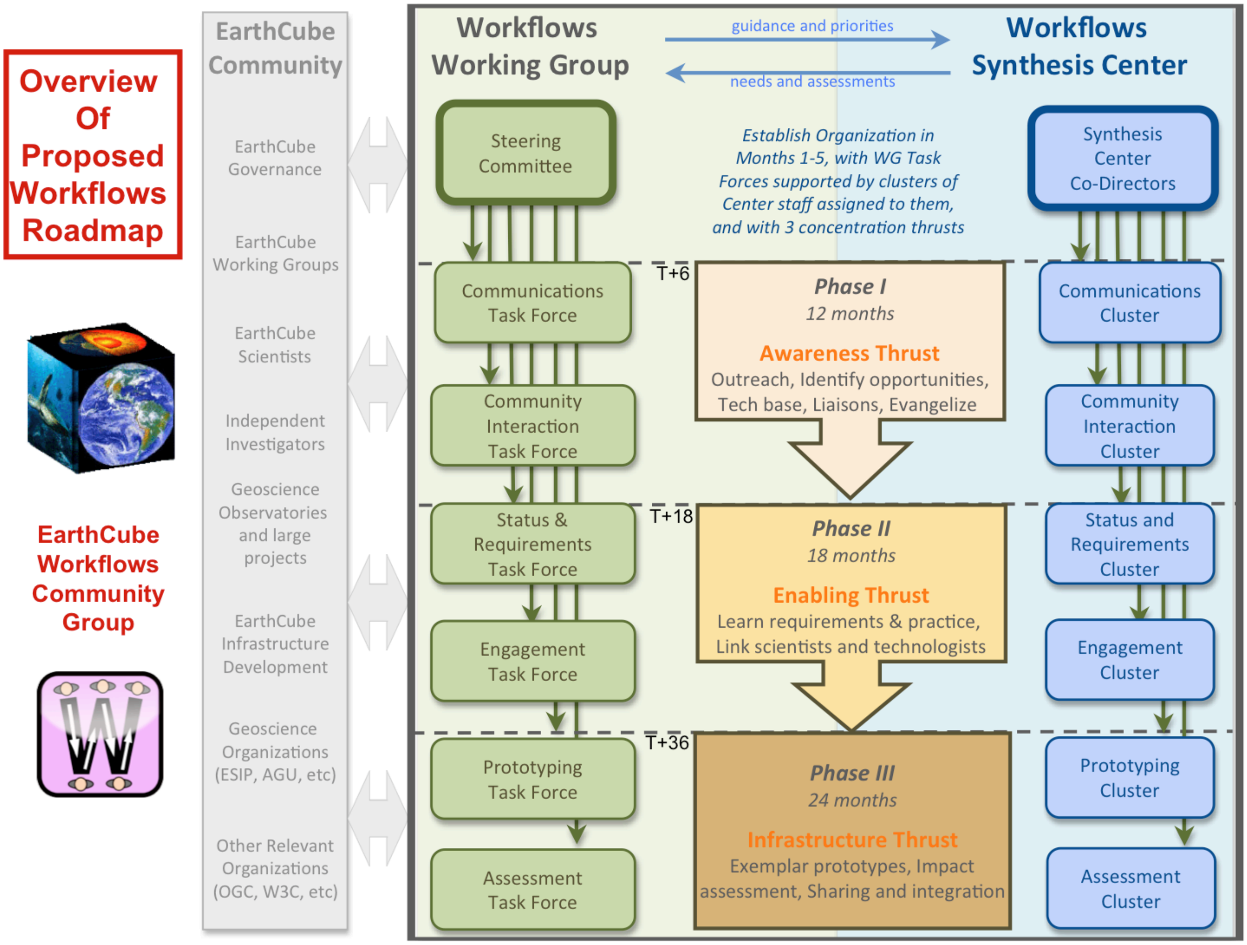
As indicated in the vignettes and documented in community workshops and surveys, geoscientists have at their disposal a wide range of approaches to workflows, but the underlying technologies used to execute them vary greatly. The primary improvements to the state of the art in the geosciences will be to increase the usage of workflow software for those groups not currently using any and to improve existing workflow systems so that they meet the needs of geoscientists (such as increasing ease of use, capturing more semantic information, and improving integration with geoscientists' resources). For these improvements to take place on a large scale, we must provide a governance model (Section 9, "Management") that will replace the "scientist as customer" model current in cyberinfrastructure development with the "scientist as stakeholder" model.

Some of the benefits of workflows for the geoscience research are (a) automation that enables the scientists to concentrate on the analysis definition and not its execution; (b) the ability to document the provenance of data that results from scientific processes; (c) providing a framework to define common formats or standards that promotes exchange of data, products, and models; (d) promote broader collaborations by multi-disciplinary workflows; and (e) ease of use - giving non-developers access to advanced models and avoiding need to download-install-learn how to use other researchers' models and codes.

Through this roadmap, we are aiming to illustrate the importance of scientific workflows in support of geoscience research, and discuss a comprehensive path towards achieving the goals and practical benefits of leveraging scientific workflows in the state-of-the-art of geoscience research and discovery.

1.6 Concrete Workshop Outcomes

The main result of this activity is a Roadmap for Geoscience Workflows. This is a living document that is publicly shared._The activities leading to the roadmap will result in many additional public documents, including a web site, presentations, shared edited documents on topics of interest, and other reference materials that are created during this activity. A graphical overview of the Workflows Community Group Roadmap is shown below. The Workflows Working Group would be led by a Steering Committee and formed by several Task Forces that are described in detail in the roadmap document. The Workflows Working Group would be supported by a Workflows Synthesis Center. The overall timeline is highlighted vertically in the middle, with the Workflows Working Group and the Workflows Synthesis Center interacting synergistically to support the roadmap activities.


Overview Of Proposed Workflows Roadmap
EarthCube Workflows Community Group
EarthCube Community
EarthCube Governance
EarthCube Working Groups
EarthCube Scientists
Independent Investigators
Geoscience Observatories and large projects
EarthCube Infrastructure Development
Geoscience Organizations (ESIP, AGU, etc)
Other Relevant Organizations (OGC, W3C, etc)
Workflows Working Group
guidance and priorities
needs and assessments
Workflows Synthesis Center
Steering Committee
Establish Organization in Months 1-5, with WG Task Forces supported by clusters of Center staff assigned to them, and with 3 concentration thrusts
Synthesis Center Co-Directors
T+6
Communications Task Force
Phase I
12 months
Awareness Thrust
Outreach, Identify opportunities, Tech base, Liaisons, Evangelize
Communications Cluster
Community Interaction Task Force
Community Interaction Cluster
T+18
Status & Requirements Task Force
Phase II
18 months
Enabling Thrust
Learn requirements & practice, Link scientists and technologists
Status and Requirements Cluster
Engagement Task Force
Engagement Cluster
T+36
Prototyping Task Force
Phase III
24 months
Infrastructure Thrust
Exemplar prototypes, Impact assessment, Sharing and integration
Prototyping Cluster
Assessment Task Force
Assessment Cluster

Additional documents prepared for the roadmap will remain active and will be immediately useful to support continuing EarthCube activities. The include the Workflow Primer, Workflow Glossary, Workflow Use Paradigms, and Workflow Vignettes.

Another important outcome is the establishment of a Workflows Community Group around workflows for geosciences that has set the seeds for future activities. This includes the formation of initial proto-task forces that have carried out various functions that have been identified in the roadmap for the Workflows Working Group Task Forces.

This Workflows Community Group now expects activities to follow the approach that has been followed so far and includes:

- Open communications: all communications occur on neutral mailing lists that anyone can join.
- Open documents: all documents are kept online and are available through the Workflow Working Group's web site. Documents can be edited collaboratively online, so they will always be living.

- Frequent, open meetings: teleconferences are openly announced and anyone can join. Meetings are vital to increase participation.

The Workflows Community Group continues to participate in community events to disseminate the results of its initial activities and the resulting EarthCube Workflows Roadmap.

References

Ewa Deelman, Gurmeet Singh, Mei-Hui Su, James Blythe, Yolanda Gil, Carl Kesselman, Gaurang Mehta, Karan Vahi, G. Bruce Berriman, John Good, Anastasia Laity, Joseph C. Jacob, Daniel S. Katz. Pegasus: a Framework for Mapping Complex Scientific Workflows onto Distributed Systems, Scientific Programming Journal, Vol 13(3), 2005, Pages 219-237.

Ewa Deelman, Yolanda Gil. Managing Large-Scale Scientific Workflows in Distributed Environments: Experiences and Challenges, Workflows in e-Science, e-Science 2006, Amsterdam, December 4-6, 2006.

Gil, Y. From Data to Knowledge to Discoveries: Artificial Intelligence and Scientific Workflows. 2009. Scientific Programming, 17(3):231-246.

Section 2. Communication

This section includes a description of a communications plan with end users, developers, and sponsors, as well as links to and feedback from other EarthCube community groups and EarthCube concept projects to promote systems integration and accelerate development. It includes a discussion of needed interactions with allied fields, agencies, and other related activities (present and desired).

Effective communication plans and mechanisms will be essential for meeting the workflow roadmap goal for ubiquitous adoption of workflow technologies. Current communication shortcomings include lack of awareness of workflow technologies by geoscientists and lack of understanding of geoscience requirements by workflow researchers and developers. This leads to problems such as invention of redundant, individually unsustainable tools and lost opportunities to collaborate on long-term challenges such as scientific reproducibility and operational efficiency. For the roadmap to be successful, we must first solve the communication barriers between geoscientists and cyberinfrastructure researchers.

A strong, well-implemented governance model is the cornerstone of the EarthCube Workflow Working Group's communications plan (see Section 9, "Governance"). The Workflow Working Group's Steering Committee is the central organizer and broker of communications. In its initial phase, the Steering Committee members were invited by the NSF to bootstrap the community organization. In the next phase, this Steering Committee must expand to include new members, both funded and unfunded, who will be stakeholders in the working group. The Steering Committee will thus need to include representative developers, end users, representatives of allied organizations, liaisons with other EarthCube working groups, and funding agency representatives.

> The Workflow Roadmap will include a governance plan for expanding Steering Committee membership to include representative groups. A broadly constituted Steering Committee with active participants from all stakeholders is essential for effective communication.

2.1 Communications Plan

The EarthCube Workflow group has practiced an open and inclusive communications policy since its inception, serving as the prototype for communications during the execution of the roadmap.

Communication Infrastructure: The communication plan's core principles include the following infrastructure elements:

1. A steering committee that spans diverse areas of expertise, geoscience focus, career stage, geography, etc. The current committee members are listed at https://sites.google.com/site/earthcubeworkflow/organizers

2. A public web site that exposes events, notes, presentations, liaisons, and all group activities such that anyone has access all the materials available. The site is at https://sites.google.com/site/earthcubeworkflow/
3. A roadmap document that is publicly readable since its first draft, open to anyone for edits, and set up as a "living document" that can be updated over time. The roadmap is at https://sites.google.com/site/earthcubeworkflow/earthcube-workflows-roadmap
4. Mailing lists that are open to the community to subscribe as well as to inspect for historical reference.

We have found this to be an effective way to help our group members keep track of our activities, help newcomers get involved, and communicate progress to other EarthCube groups and to the NSF. We will continue to practice this open philosophy for communication with the community.

Communication Activities: It its initial phase, the Steering Committee has engaged the geoscience community through a number of venues to encourage open discussions that should serve as a plan for future, continued engagement:

1. Outreach to the geosciences community through established venues. Activities initiated during the Spring/Summer 2012 include:
   a. An EOS Science Brief, *"Designing a Roadmap for Geoscience Workflows"*, published in 12 June 2012, as an initial publication of the geoscience workflow concept and invitation to the community to participate.
   b. A session at the 2012 ESIP Summer meeting where the roadmap will be discussed with the community
   c. A session at the American Geophysical Union's Fall 2012 meeting.
2. Creation of introductory materials to describe workflow technologies in ways that are accessible for geosciences researchers that are not familiar with this area. These materials continue to be improved based on community feedback. We believe these materials are already helping in starting a dialogue with geosciences researchers that have never heard of workflows. They are publicly available at https://sites.google.com/site/earthcubeworkflow/about-workflows. These include:
   a. a workflow primer document,
   b. a set of prototypical workflow use paradigms,
   c. a glossary of common terms, and
   d. a series of vignettes about current or potential uses of workflows.
3. Creation of a Workflow Community Questionnaire to convey what aspects of geosciences infrastructure are of particular interest to our group. Answers to this questionnaire are collected on-line at https://sites.google.com/site/earthcubeworkflow/questionnaire-for-the-community
4. Conducting regular virtual meetings with clear goals, agendas that are inclusive for participants, and with immediate publication of slides and notes from the meeting sessions.
5. Choice of workshop physical location at a site that would draw most participation from the geosciences community. Our only face-to-face workshop to date was held in

Boulder, CO, and included participants from many local organizations as well as remote participants.

6. Approaching all the other EarthCube groups and interacting with them one on one. These activities are described in more detail below.

The Workflows Roadmap will include a Communications Task Force that will be responsible for both communication infrastructure and events. The Communication Task Force will disseminate information to the community through: 1) a web presence with accessible (non-technical) information for geosciences researchers, 2) clear points of entry and participation for engaging newcomers, 3) pursue synergistic and opportunistic meetings of for both virtual and physical participation, and 4) measure the growth of the participating community and their engagement over time.

Communications with End Users

The goal of our communications plan is for a wide range of geoscientist to be first-class participants and stakeholders in the EarthCube Workflow Working Group, effectively doing away with the concept of "end users". Already many of our activities are designed to engage the broad community, such as our questionnaire, our collection of vignettes, and our introductory materials.

We will continue to proactively engage the geoscientific community by pursuing opportunities such as the following:

- We wrote an EOS note that introduced the activities of the EarthCube Workflows group and communicated our interest to receive inputs from the community. The note was accepted and will appear soon. EOS has broad dissemination in the geosciences community. EOS Science Brief, "Designing a Roadmap for Geoscience Workflows", to be published 12 June 2012.
- We will be holding a session at the ESIP Summer 2012 meeting
- We have proposed a session at the 2012 AGU Fall Meeting

We also plan to actively engage the upcoming EarthCube workshop awards that will be targeting various areas of geosciences. Our questionnaire will provide a structured framework for guiding discussions. The submission of vignettes will be another mechanism for end users to communicate their interest in workflow technologies. We will work with NSF and the workshop organizers to develop appropriate mechanisms to engage workshop participants and their broader communities.

The Workflows Roadmap will include a Communications Task Force that will organize and actively participate in community-wide EarthCube workshops and other significant community events by 1) defining goals and expected products from a workflows session at the workshop, 2) preparing materials to facilitate engagement from workshop participants, 3) establish strategic lines of communication with that community that are sustainable in the longer term.

The Workflows Roadmap will include a Communications Task Force that will pursue publications through 1) submitting regular reports in newsletters and journals regarding group activities and progress, 2) encouraging group participants and the community to publish significant accomplishments, 3) collecting published materials and making them publicly available.

A key aspect of engaging the geoscientific community is to demonstrate the benefits of workflow technologies for doing science. Therefore, a key component of our roadmap is to pursue the development of prototypes of workflow technologies that can demonstrate key capabilities. This will help cyberinfrastructure researchers and workflow developers gain advocates among scientists, which will help in evangelizing and disseminating workflow technologies. This will also help all participants to understand difficulties with adoption as well as standing socio-technical challenges that will require further research in workflows. Workflows can help students harness complex models and science grade data in an accessible manner, with great potential to impact education and cross-disciplinary research. These prototypes can grow to become operational systems, and hopefully demonstrate the benefits to scientists in terms of accelerating the pace of their research and making discoveries that they would not otherwise even consider pursuing.

The Workflows Roadmap will include a Prototyping Task Force to pursue prototypes of workflow technologies in various domains of geosciences can: 1) provide exemplary demonstrations of the capabilities of workflow technologies for doing science, 2) gain advocates among scientists who benefit from workflow technologies, 3) uncover difficulties in adoption and need for outreach activities, 4) disseminate expertise to manage this novel aspect of cyberinfrastructure, 5) create opportunities for interactive educational materials for students in their geosciences domain of interest, and 6) produce scientific advances and discoveries that would not occur as fast or as easily without workflows.

Communications with Developers

Developers and end users of workflows should be thought of as existing on a continuous spectrum. There will some who will be satisfied with taking a workflow tool and using it as is; some who will be active developers of their own applications, services, etc, and who will understand the workflow tool in detail and be able to give detailed feature requests, contribute modules, etc; and some who will actively contribute to the core software of a particular workflow tool.

The Workflows Roadmap will include a Communications Task Force that will organize and participate in events attended by cyberinfrastructure developers. Examples include AGU, XSEDE annual conferences, and Supercomputing.

Based on our team's experience, we will engage workflow tool developers and others involved in cyberinfrastructure research through open community processes. An important shortcoming of many cyberinfrastructure projects is the reliance on standards without an open source, open community-based reference implementation. That is, there is not a neutral development venue with well-defined governance processes for implementing the standards. The Apache Software Foundation provides an operational model for open communities, with a well-defined governance model (the meritocracy) and rewards for community diversity. GitHub, SourceForge, and Google Code are less formal models: open communities can build on top of these services, although no specific governance model is provided.

The advantage of this model is that it converts users into stakeholders by rewarding contributions to a project with full membership in that project. For the EarthCube Workflow Working Group to succeed, it must attract and reward unfunded developer participation in its various activities.

Participation in standards-developing organizations will remain important, particularly W3C and OGC.

The Workflows Roadmap will include the establishment of an EarthCube Synthesis Center on Workflows that will 1) include senior researchers, post-doctoral students, and research programmers with expertise in workflow technologies with some geosciences background, 2) conduct courses for students about workflow research and for developers about workflow software adoption, 3) host visitors that will engage in prototyping activities with center personnel.

Communication with Sponsors

A great potential value of the Workflows Community Group and the Roadmap activity will be to identify and communicate research priorities and opportunities and grand challenges to the NSF and other sponsors. These recommendations may result, for example, in modifications to planned funding solicitations to address priorities identified by Workflow Roadmap requirements processes or even in the creation of new solicitations.

The Workflows Roadmap will recommend that the Community Interaction Task Force make regular (annual) recommendations to the NSF concerning opportunities and unmet needs obtained through requirements gathering and related activities.

More mundane communication processes of the Workflows Community Group with the NSF and other sponsors will include reports regarding progress, success stories, and challenges in its activities.   We will quantify the technical and scientific outcomes of our activity, convey the breadth of our outreach, and report on its impact in education. We will develop periodic reports of priority areas of investment in support of our roadmap as it evolves.

2.2 Integration with Other Community Groups and Concept Awards

During its initial phase, the Workflows Community Group has established communication channels with the other community groups and concept awardees. This serves as a prototype for future communication plans, which will need to continue throughout the EarthCube project.

EarthCube Community Groups

Preliminary Activities: There is a clear need to maintain regular interactions between the Workflows Community Group and the other community groups. The Workflow Community Group has established several liaisons with other groups that will serve as points of contact and are listed on the group's web site:

- Governance: Shahani Weerawarana, Indiana University
- Semantics and Ontologies: Nicong Li, University of Redlands
- Data Discovery, Mining, and Access: Suresh Marru, Indiana University

From early on in the project, we have had interactions with the Governance group. We had an invited discussion with Lee Allison, lead of the Governance group, and members of both groups. One important benefit for our group is the existence of the Governance Forum, a group with 60 members that represent broadly the EoIs submitted to EarthCube. We have used the Governance Forum as a source of points of contact for a variety of groups in the community that we have started to contact (see section 1.3). We need to continue the conversation to convey the needs of the Workflows group regarding an appropriate governance, and reflect those as EarthCube organizations emerge.

We had a special session at one of our weekly meetings with the leads of the other Concept groups. The session included Krishna Sinha from the Semantics and Ontologies group, Rahul Ramachandran from the Data Mining group, Tanu Malik from the Data Access Services group, and Chaitanya Baru from the Data Discovery Services group.

The Semantics and Ontologies group has clear relevance to the workflows group. Ontologies can be used to formalize large-scale, cross-disciplinary planning process workflows, domain process workflows, as well as scientific workflows. Besides coding the various levels of workflow templates, ontologies can be used to semantically annotate the resources needed for instantiating a workflow template (data sets, models and tools) by, for example, indicating their purpose or classification. Semantic registration of these resources is essential for automatic resource discovery on CyberInfrastructure, which is itself essential for automatic workflow orchestration. Coupled with semantic reasoning, ontologies can further guide workflow template instantiation, or guide new workflow template composition.

There are many ways that the attributes of a specific planning process can semantically inform lower level workflows. For example the type and characteristics of a specific planning problem (e.g. site search or selection) may provide guidance or constraints on the type of algorithms (e.g. optimization) to be used in the computational workflow during the solution alternative design step, or knowing that the domain is species recovery for a desert dwelling burrowing species (desert tortoise) could constrain the type of hydrological models employed to rainfall runoff models. Similarly, when composing a scientific workflow under a domain process computational workflow, the semantic information on the "entity type" being considered and the bounding geographic area for the entity distribution in the domain process workflow can be used to specify the input data requirement for the scientific workflow. Some of these semantic constraints are propagated down from the planning process workflow (e.g. from the planning objectives and planning spatial extent) to the domain process model workflows. All this will furthermore affect the choice of software tools (which implement algorithms) to be used during a scientific workflow.

Effective collaborative research between the Semantics and Ontologies Working Group and the Workflow Working Group on workflow research is essential. This has been recognized during the Semantics and Ontologies Workshop (April 30 - May 1) as well as during the Workflow Workshop (May 16). Collaboration activities thus far include representatives of one working group participating in the workshop and weekly meetings of another working group (e.g. by Yolanda Gil, Pedro Szekely and Naicong Li), in an effort of identifying the common research topics between the two working groups, briefing each other of current research activities, and sharing new ideas. The Technology Committee of the Semantics Working Group also drafted a report including sections on workflow related topics. It has been suggested that collaboration between these two groups can further benefit from working on a common Grand Challenge type problem use case. Large-scale environmental planning problems can provide such use cases since they involve applying computational workflows to process massive amounts of heterogeneous spatial data with ever increasing analytic complexity, work that cuts across different earth science domains to social and information sciences, and can provide end-to-end interoperability use cases for EarthCube initiatives.

The Data Discovery, Mining, and Access group has many aspects that are closely related to workflows. Workflows are often used for data mining. Workflows also express declaratively data requirements that could be used in combination with data discovery capabilities. Workflows access data dynamically during execution, and generate data products that often need to be shared and made accessible by others.

Roadmap Activities: We anticipate these discussions to continue. Specific themes could be pursued at community workshops and meetings to flesh out the potential common interests. For example, at the Workflows May 16 2012 workshop, a breakout session on Semantics and Workflows was formed (notes are available on our group’s site). Recurring themes could form more official Special Interest Groups (SIGs) within EarthCube, strengthening the role currently held by the liaisons across groups.

Concept Projects

Preliminary Activities: The prototype development activities and longer term visions represented by the concept projects are directly relevant to the Workflows group. We have established several liaisons with each of the projects:

- Brokering: Steven F. Browdy, OMS Tech, Inc.
- Interoperability: Ilkay Altintas, UC San Diego
- Layered Architecture: Ilkay Altintas, UC San Diego
- Web Services: David Fulker, OPENDaP
- Earth Systems Modeling: Scott Peckham, University of Colorado

We have engaged all these groups through participation in our virtual workshops and weekly meetings, as well as participating in their discussions.

The Brokering project has put forward a design with many brokering services supporting workflow components in the architecture. They have also designed several forward-looking use cases where we could identify slices where workflows would be useful. For example, they have developed a use case for volcanic eruption analysis (disruption of air traffic, environmental impact, etc), where the vignette for air turbulence simulation would be relevant. Another use case concerns reduction of hypoxic waters in the Gulf of Mexico (from land use and other pollutants from the Mississippi River), where the vignettes for water quality and flooding models would be relevant. Further exploration of these use cases would lead to many more opportunities for use of workflows that would be combined with brokering capabilities for data access, semantics, and other services.

The Layered Architecture project has identified workflow environments as one of the layers within a higher level collaboration environment. The group seeks to define the mechanisms for instantiating workflow systems within collaboration environments to perform different steps in the end-to-end process for a scientific study, e.g., access to data repositories and coupling visualization tools to accessed datasets. We are working with a liaison from this project to identify scenarios for making workflows a part of a researchers' daily work tools. A typical such scenario consists of a researcher acquiring data from a community data repository, the processing of the data by an analysis workflow created by the researcher, and the storing of the results back to the repository within the research environment.

The Interop project has identified workflow scenarios to identify requirements via workflow scenarios that span across scientific domains. These workflows help identify cross-domain issues including data and file formats formats, units and temporal and spatial representations. Workflow scenarios include atmospheric forecast models, coupled atmospheric/ocean models, coupled atmospheric/hydrological models, and coupled atmospheric/hydrology/storm surge models. We established a working relationship with the Interop project through a project liaison to learn workflow modeling requirements related to such cross-domain scenarios. These requirements are summarized as challenges (section 3) in the roadmap for the Interop project and include ensuring metadata and content in the cross-domain context and working with different encoding formats for different analysis and display tools. For example, the hypoxia use case integrates hypoxia data from several disciplines (terrestrial hydrology, physical and biological oceanography, and potentially meteorology and satellite sensing) as well as data with different encoding formats, scales, and sampling geometries. The developed workflows had to match the differences in the data formats and solve challenges on data discoverability, accesibility and downloadability that is not standardized.

Promoting Systems Integration

Our Workflow Use Paradigms can be useful to detect possible opportunities to integrate workflows in use cases and prototyping efforts pursued by other groups. The workflow vignettes have also proven useful to provide connections across vision scenarios and potential uses of workflow technologies.

The integration activities pursued will be driven by interesting demonstrations of workflow technologies and/or science value.

2.3 Future Interactions with Other Activities, Fields, and Agencies

We initiated discussions with a number of individuals representing large projects and communities. Our starting point was the Governance Forum of the Governance group. We engaged many organizations through invitations to speak at our virtual workshops as well as weekly meetings. The communities and individuals that we have interacted with so far include:

- Major funded projects:
    - Brian Wee on NEON (National Ecological Observatory Network)
    - Anthony K. Aufdenkampe on CZO (National Critical Zone Observatories)
    - Richard Hooper on CUAHSI (Consortium of Universities for the Advancement of Hydrologic Science)
    - Matt Jones on DataOne

- Major standards efforts:
    - Dave Fulker on the OPeNDAP organization (Open source Project for a Network Data Access Protocol)
    - George Percivall and David Arctur on OGC

- Major government agencies:
    - Ben Wheeler on USGS
    - Chris Mattmann and Paul Ramirez on NASA/JPL

- Major community organizations:

    - Erin Robinson on ESIP
    - Gerry Wiener on UCAR/NCAR

There are many more projects and organizations in geosciences that could be engaged regarding a workflows roadmap. We have found benefits through the interactions with these communities to date, including:

1. existing use cases that have been articulated and in some cases pursued by the community.
2. working groups and activities that are already using workflow technologies and can provide success stories, lessons learned, and future requirements and challenges
3. opportunities for events that would enable us to do community outreach.

Specific use cases and activities discussed to date include:

- DataOne has developed both science-driven and infrastructure-centered use cases that will be useful for our group's activities:
    - The Exploration, VIsualization, and Analysis (EVA) working group is currently engaged in a project for intermodel comparisons for carbon-climate model output. They use workflow technologies and of particular interest is the combination of visualization, workflow, and provenance. A previous project by this group was on continental scale bird migration, also using workflow technologies.
    - There are also very detailed use cases of how different infrastructure capabilities interact in the DataOne federation. Some use cases are directly relevant to workflows.
- ESIP has regular open meetings where we can engage the community regarding workflows. Activities include:
    - The ESIP Summer 2012 meeting had a session devoted to the EarthCube Workflows Community Group on July 18, 2012 that will be co-chaired by Chris Mattmann and Paul Ramirez from NASA/JPL and Yolanda Gil of USC/ISI. The goals will involve audience participation in discussing the EarthCube Workflows roadmap, collection of additional vignettes, planning for the next meeting, and collection of input to the NSF.
    - The ESIP Earth Science Collaboratory has developed several scenarios that include workflow technologies. These scenarios illustrate the integration of various capabilities relevant to EarthCube.
- OGC has a Workflows Working Group as well as other activities relevant to workflows such as:
    - OWS-6 Geoprocessing Workflow Architecture: This Engineering Report provides a summary of Geoprocessing Workflow practices and methods in a SOA environment.
    - OGC Web Processing Service: The WPS Standard provides rules for standardizing how inputs and outputs (requests and responses) for geospatial processing services. The standard also defines how a client can request the execution of a

process, and how the output from the process is handled. It defines an interface that facilitates the publishing of geospatial processes and clients' discovery of and binding to those processes. Some workflow systems have been used to demonstrate these WPS services.

○ OGC Workflow Domain Working Group: The mission of the Workflow DWG is to establish a forum for describing, discussing, and solving any issues related to geospatial workflows. By geospatial workflow we mean any workflows that contain any or all processes that relate to geospatial processes and/or data. The primary focus of this DWG is to help individuals and organizations to identify smarter, easier, and more economical ways to build, migrate, manage, and maintain workflows.

The Workflows Roadmap will include a Community Interaction Task Force to pursue and coordinate interactions with major funded projects, standards efforts, government agencies, and community groups in geosciences to: 1) uncover use cases and opportunities for dissemination of workflow technologies, 2) collect success stories, lessons learned, and future requirements and challenges for workflows in geosciences, 3) leverage their efforts by aligning them with the Workflows roadmap, and 4) take advantage of opportunities to participate in events where large communities already participate.

The Long Tail of Geoscience: The Long Tail of geoscience research represents a broad community of researchers that are not yet being served by best practices for workflows that support fundamental access to geo-data, models and collaboration resources in pursuit of science. There are many reasons for this but one major issue is the inefficient or limited access to Earth Science Data necessary for their research. For example the basic data for soils, surface geology, climate reanalysis, vegetation/land cover, aquifers, streams, and terrain all reside on on separate federal servers with relatively slow access and in formats foreign to most Long Tail scientists and their students. It can take months to assemble all the necessary basic data to conduct a research effort. The lack of access to fundamental Earth data delays the research effort and seriously effects the added value of new experimental data performed by researchers. This lack of efficient access to national data limits the process of versioning and provisioning basic data for improving the quality and resolution of soils, hydrogeology, land cover, terrain, climate data, etc. Given the rate of new data, model and model-data products that are being developed by agencies and researchers, it is imperative that workflows be designed and implemented that improve access to fundamental Earth science data and the models they support.

Interactions with Other Fields

Interactions with other fields would be very beneficial.

Computer and Information Sciences and Engineering: Basic research on workflows occurs mostly in Computer Science, and through these interactions there will be a faster flow of new techniques into geosciences. It is also important to convey to computer science researchers the future challenges and opportunities offered by geosciences for advancing the state of the art in workflow research. The requirements for workflows in geosciences will not be achievable without the active involvement of this research community.

Cyberinfrastructure: The NSF Office of Cyberinfrastructure funds a significant amount of workflow software development and deployment through its SDCI, SI2, and STCI awards. SI2, in particular, has an organizational component, with a recent workshop (https://wiki.ncsa.illinois.edu/display/SI2/Building+Communities+for+SI2). The planning grant awards for large scale (S2I2) awards within this program should als be announced soon. The Workflows Community Group should establish liaisons with appropriate awardees and organizational efforts. Finally, there are a number of large scale cyberinfrastructure efforts, including XSEDE, the Open Science Grid, and domain-centric efforts such as iPlant that the Workflows Community should interact with. These efforts already have well-defined workflow efforts, so matching these to geoscience users and research groups will be important.

Biomedical informatics: Workflow systems are beginning to be used in a variety of areas in bioinformatics, notably genomics and image analysis. Exchanging lessons learned regarding technologies and infrastructure would be very beneficial.

Informatics in various disciplines: This includes activities with ecoinformatics, geoinformatics, etc. in the community.

The Workflows Roadmap will include a Community Interaction Task Force that will engage computer science researchers with EarthCube activities will benefit the geosciences community by enabling: 1) defining basic research needed in workflows motivated by grand challenges in geosciences, and 2) faster transfer of new advances in basic workflow research into geoscience infrastructure.

Section 3. Challenges

This section includes a description of major drivers, trends, and shifts impacting or that could impact the focus of a working group, including but not limited to changing technology, adoption culture, and community engagement.

The overall goal of the Workflow Roadmap will be to make workflows ubiquitous within the geosciences. This section reviews challenges derived from this goal.

3.1 Challenges to Adoption

While there are a number of workflow systems that are used and/or well-known in the geosciences community, we found through our survey that some groups are developing their own workflow technologies from scratch. The tension between encouraging adoption of mature workflow systems versus development of lightweight customized systems or simple scripting solutions will need to be addressed.

A large percentage of geoscientists are not using any workflow tool. This has a number of consequences: lost efficiency, lost metadata, lack of reproducibility, limited or no access to national geoscience datasets, problem of national geo-spatial/temporal data on secure federal servers with many different formats, etc. The challenge is to increase the access and efficiency of access to geoscientists.

The Workflows Roadmap will include an Engagement Task Force that will: 1) develop guidance for scientists in finding appropriate approaches and systems to address their workflow needs; 2) assist scientists in evaluating potential workflow technology solutions; 3) request the support of other task forces, i.e., the Status and Requirements Task Force and the Prototyping Task Force, when necessary; and 4) disseminate their expertise in workflow solution approaches.

3.2 Challenges in Addressing Requirements for Workflow Technology

Reproducibility, the corner stone of the scientific method was indentified as an important problem by the respondents of the survey we put forward to the community. Reproducibility was important whether it involved individual scientists or groups producing standard data products. True bit-by-bit reproducibility may be an impossible problem (particularly for workflows involving high performance computing, which at least depends on the underlying applications to have invested a lot of effort into it). We can instead consider this to be a problem of asymptotically approaching reproducibility.

The Workflow Roadmap will include a Status and Requirements Task Force that will investigate reproducibility issues in the geosciences.

There are many challenges stemming from the presence of a number of workflow management systems that possess various capabilities and features [1]. DIfferent workflow systems have started out by targeting various user communities. Some, like Kepler, have focused on providing rich graphical interfaces for easy workflow composition by individual scientists. Some, like Wings, have included semantic component and data descriptions to guide the user in their workflow-building tasks. Other workflow systems have focused primarily on issues of related to workflow execution reliability and scalability (DAGMan, Pegasus) and on applying service oriented architecture principles to workflows on widely distributed resources (Apache Airavata).

Just as scientists use a variety of programming languages (FORTRAN, C, JAVA, python), so do they use different types of workflow management systems.

The Workflow Roadmap will include a Status and Requirements Task Force that will identify the classes of workflow tools, provide example tools, and determine gaps..

However, unlike in the language programming space, there are many more workflow management systems being developed today. Some are geared towards specific communities, or even research group. In some sense it is fairly easy to develop a program that sequences a number of user-defined steps. Only with time do the software builders realize the complexities of the overall problem (portability, ease of use, robustness, scalability, etc) and the challenges in providing and supporting software solutions in the long run.

As with many software capabilities, the sustainability of workflow systems over time is a challenge for the earth science and other communities. Single-author tools may play an important role, but what happens to the projects that depend on this tool when the main developer is no longer involved? The software needs a way to find a broader collection of stakeholders. At the other end of the spectrum, depending on tools by a large commercial vendor also carries risks. The community needs a mechanism to make sure that the vendors support their software, donate it to an appropriate open community venue if the software is discontinued, and similar actions.

Another challenge is workflow interoperability. The need for interoperability would likely arise when workflow systems are more widely used and scientists would want to move their workflows to another system. Suppose a group of scientists uses a particular workflow system to develop a workflow. Months later, they realize they need a capability that the workflow system in question does not address very well, and they wish to try another one (or several other ones). How would this be supported? Some interoperability mechanism should be provided so they would not have to start from scratch in the new system.

The big question is, at what level should the workflow systems inter-operate. A European project, SHIWA (http://www.shiwa-workflow.eu/) is examining such issues. It defined two levels of interoperability: coarse and fine-grained. Coarse-grained interoperability enables the output of one workflow to be fed into the input to another workflow managed by a different system. A tighter interoperability is implied by the fined grained approach, where two workflow management systems interoperate via a common intermediate workflow representation. Currently, there are no interoperability efforts in the US, and the question remains as to what impact these efforts can have on the productivity of the scientists.

Another grand challenge topic is the capability of workflows to utilize data coming from diverse communities and disciplines. Workflows should not have the responsibility of dealing with this type of interoperability. It is, arguably, not a core competency of workflows. Brokering frameworks in the geosciences are beginning to emerge and evolve that do provide this capability, and in some cases with semantic mediation for vocabularies, what some would call "light semantics." This brokering framework, in the cyberinfrastructure, would be able to discover and access data needed by workflows, and provide it to the workflows in the standard that the workflow requires. This works based on well-known interoperability standards (protocols and formats) that allow brokering to access source data, convert it to what a workflow expects, and provide it to the workflow.

Finally as the scale of the data sets and of the computation grows, and the diversity of the data increases (lidar, climate reanalysis, vegetation, etc.) there are very important challenges that are presented to the workflow management systems. Examples include how to handle interactivity, checkpointing, debug modes, layering of workflow capabilities (composition/ execution) error propagation, reliability, deep communications with underlying resources (codes and libraries, computing nodes, etc).

Since analyses are being composed not of only one workflow but many workflows (ensembles) it is necessary to manage these workflows as collections, examining issues of scheduling, resource sharing, fairness, time to solution, etc...

3.3 Impacts of Changing Technology

Many technologies relevant to workflows are changing:

1. Computers, clusters, supercomputers with many different architectures. Codes have to be ported. Reproducibility is a challenge.
2. Commercial clouds and their offerings.
3. Size of many data products and the difficulty of moving them around.
4. Proliferation of streaming data from sensors.
5. Need to share, discover, and control data, other social networking issues.
6. Need to develop models that incorporate multiple data types, and also the need to know when it is ok to leave data out from your model.
7. Mobile devices, ubiquity.
8. Levels of funding available.
9. Brokering frameworks that facilitate cross-disciplinary data sharing.
10. Data intensive computation in geosciences often requires fast access to data during model runtime. Standard computer clusters typically do not incorporate the data-intensive aspects of geoscience models. New technologies with fast storage on multicore processors may help resolve this problem.

These and other technologies relevant to workflows are likely to be changing, in addition to new technologies that we cannot imagine yet.

Changing technology is always a challenge for scientists. Yes, they need to keep up with the new developments that impact their work, however, they need to see a real return on the investment of their time and effort to learn new tools. Scientists that are part of larger projects and that use the national cyberinfrastructure and who require scalability, or a large degree of automation, often have access to IT people to help them port their computations to new technologies and infrastructure. The challenge is to provide the same productivity gains to the "long-tail of science" researchers, so that they can benefit from the new technologies as well.

The changing technology and infrastructure also have an impact on the workflow management system developments. These systems need to be maintained and enhanced over time so that they be relevant and useful to science as the underlying technologies evolve.

3.4 Impacts of Community Engagement

The idea of using workflows is not widely known to geoscientists. A very small fraction of the community uses workflows. Yet, many science processes could be managed with workflows. Data preparation processes often involve many steps that could be managed by a workflow system. Data analysis processes are very common and are often multi-step and complex in nature, and could be managed with workflows.

Currently, data is typical currency in geosciences, as is the case in many other sciences. Models are also considered important but to a lesser degree. For example, while many data catalogs and data management services exist, there is very little work in the community in terms of model catalogs or model-related services.

Essentially, all scientists care about data, many care about models, and very few care about workflows. This forms a pyramid of sorts and emphasizes the gravitas that data takes in science. A good goal for the Workflows Working Group is to change the shape of this pyramid to reflect broader interest in workflows.

However, there is a cost to learning to use a workflow system, and the benefits must clearly outweigh the cost. Only if we articulate the benefits clearly, and they are significant, will there be uptake in the community.

Therefore, a challenge for a Workflows Working Group will be to educate the community on workflow technologies, while understanding the cost and benefits of adoption.

The group should synthesize data from current workflow users regarding cost/benefit analysis. Research focused on understanding cost/benefit analyses of workflows will need to be carried out.

References

[1] Ewa Deelman, Dennis Gannon, Matthew Shields, Ian Taylor, Workflows and e-Science: An overview of workflow system features and capabilities, Future Generation Computer Systems, July 10th 2008.

Section 4. Requirements

This section includes a description of process(es) to be used to get the necessary technical, conceptual, and/or community (i.e., end-user) requirements at the outset and during the life of the activity, including approaches to achieving community/end-user consensus.

4.1 Challenges and Preliminary Findings

The Workflow Roadmap will need an ongoing process for obtaining and understanding requirements of the geoscientific community. The diversity of users is an important challenge that must be addressed when obtaining these requirements.

User Diversity:

- Field researchers: these may be more interested in data collection and analysis. Their workflows may be only partially digital (ie need to make sure the field sample was properly annotated and stored).
- Computational researchers: these are researchers who are constantly building and modifying their models. Each one can be unique.
- Data-Intensive Computational Researchers: these are researchers who assimilate historical or real-time streaming data in models, or people who use other models as input (climate models as input to hydroecology models).
- Operational users: these are people who are responsible for creation of public data sets, forecasts, etc. These are probably good candidates for many current workflow tools.
- Classroom usage: Many workflow tools would expose students to analytic tools in science through workflows.
- Graduate and undergraduate research support: Geoscience student researchers spend an extraordinary amount of time attempting to find and access data necessary for their research. This can be everything from geospatial data, to time series, to publications (digital and paper), maps (digital and paper), relational databases, etc.

These may each require different types of workflow capabilities.

Preliminary Findings: As part of its March-June 2012 workshop series, the Workflows Community Group created a questionnaire (https://sites.google.com/site/earthcubeworkflow/questionnaire-for-the-community) as a way to capture community input. The survey format allowed essay responses to questions. From the community survey responses so far obtained, efficient sharing of multi-step data transformations, handling big data, projecting diverse geospatial/temporal data sets, integrating multiple data sets, managing complex executions, reproducibility of results, and interoperability with other tools and services (OPENDaP, NetCDF, OGC services, ArcGIS, etc.) are all capabilities mentioned by the responders.

Surveys will be an important mechanism for future broad requirements gathering. The initial survey was useful in making new contacts and in formulating initial conceptual requirements, but the format was not appropriate for gathering results that can be statistically analyzed. The initial questionnaire can be improved in successive versions (with help from experienced survey designers) as representative prototypes of geoscience workflows are designed, implemented, and evaluated.

The Workflows Roadmap will include a Status and Requirements Task Force that will be charged with formulating and executing plans for obtaining user, conceptual, and technical requirements.

The Workflow Roadmap will need to include methodologies for more systematic requirements gathering. The remaining sections outline these processes.

The Status and Requirements Task Force will follow well defined processes and assessment metrics to gather user requirements and to understand the impact of workflow technologies in geosciences research.

4.2 Processes for Obtaining User Requirements

We outline here a general approach to user requirements gathering. As discussed above, user diversity is a challenge for requirements gathering, so the process must take this into account. Since the user space is large, a guided study that samples important requirements space will be more useful than a brute force search.

Startup Processes

1. Broadly associate the range and types of geoscience data and model needed by scientists, grad students, K-12 teachers, environmental resource managers and the public. For example, geoscience data can mean something like “Essential Environmental Variables” (EEVS). The WMO uses the terminology Essential Climate Variables but that is too narrow. Most of the EEVS are already accessible but they reside on many different federal servers often with slow access and in formats not understood by geoscience users. (e.g. topography, historical climate or reanalysis, landuse and land cover, geology, soils, etc.).
2. Develop a matrix for aligning workflow technologies with the use-cases outlined earlier and the particular data and model needs in each use case. Partition the matrix into near-term (1-4 years) and long term 5-10 years) workflows.
3. Carry out a hypothetical prototype analysis on an existing workflows that can answer the following:
    - How can workflows automate the management and sharing of data and models, experimental design, as well as efficient sharing of software?
    - Can workflows improve the geoscientists ability to track, visualize and analyze data, experiments and models results through a unified framework?
    - What is the role workflows in tracking provenance of data and models in the context of scientific reproducibility and for advancing scientific understanding?
    - How does a workflow facilitate data and model quality assurance?
4. Conduct a community survey to evaluate the matrix and the prototype.

Follow-up Processes

1. Design a technical strategy for aligning workflow technology for each geoscience use cases outlined earlier.
2. Design 2 to 3 pilot studies from the matrix of examples of workflow technologies over the next 1-4 years (near-term roadmap) and 5-10 years (long term roadmap)
3. Present results of preliminary design at the AGU/IEEE national meetings to solicit feedback from the larger geoscience computer science communities.

Consensus Processes

1. Develop an evaluation strategy for testing the effectiveness of the 3 workflow prototypes
2. Demonstrate how the workflow improved the efficiency of the geoscience team for each workflow prototype

4.3 Processes for Obtaining Technical Requirements

Technical requirements will be obtained primarily from the cyberinfrastructure community and are derived from geoscientist community requirements (Section 4.2).

Startup Processes

1. Data: For example, develop a workflow structure for making essential terrestrial geo-data from many federal agencies accessible within the workflow (e.g. topography, historical climate or reanalysis, landuse and land cover, geology, soils, etc. see UN site http://www.fao.org/gtos/doc/pub52.pdf). This will also include agents/tools for automating data collection, transfers, data classification, data derived products and data management.
2. Models: Implement workflow technologies for geoscience models for the proposed workflow prototypes discussed above. This should include simple conceptual mathematical models developed in MATLAB or Mathematica and HPC-level models that run on large clusters or the grid. Leverage exisiting community resources such as CSDMS (e.g. Community Surface Dynamics Modeling Systems), the NCAR/NOAA Community Model Weather Research & Forecasting Model and others.
3. Fault Tolerance, Quality Assurance & Provenance: The workflow environment will require the automated capability of identifying failures in system components, generally evaluating errors in data and models, versioning of models/data and automating fail-over strategies.
4. Research Planning and Scheduling: Implement technical requirements for user-directed scheduling of workflow technologies for data and models that align with the use cases outlined earlier and data and models needs in each case. This will depend on the geoscience problem or hypotheses but we might classify them as: retrospective simulation based on the geologic past or recent climatic changes, real-time simulation for data assimilation form streaming sensors, and finally prediction or projection-type simulations (e.g NCAR/NCEP WRF or IPCC projections).
5. Research Discovery: How do the proposed workflow tools support the intersection of data, models, analysis and visualization across geoscience disciplines?

Follow-up Processes

1. Develop 3 representative prototype examples from the matrix of examples that will most likely be enriched by workflow technologies over the next 1-4 years (2 prototypes) and 5-10 years (1 prototype).
2. Implement the representative workflow prototypes with strong community interaction and participation through funding from EarthCube.
3. Evaluate the protoype workflows
4. Carry out a community survey to evaluate the matrix and the pilot design prototype.

Consensus Processes

1. Present results of preliminary design and community survey at the AGU/IEEE (?) national meeting(s) to solicit feedback from the larger geoscience and information science communities.

4.4 Processes for Obtaining Conceptual Requirements

The user requirements gathering process will most likely result in gap analysis. Conceptual requirements gathering will face the more challenging problem of identifying conceptual omissions and shortcomings in the current workflow research landscape that have implications on geoscience workflows. These requirements would typically result in recommendations to the NSF on research opportunities suitable for solicitations.

Startup Processes

1. Refine the matrix for aligning workflow technologies with the use cases outlined earlier and with the particular data and model needs in each use case. Partition the matrix into near-term (1-3 years) and long term 4-10 years) to resolve the needs of the prototype workflows that can be achieved relatively quickly (1-3 years), mid-term (4-6) and longer term (7-10).
2. Evaluate the matrix for aligning workflow technologies with the use cases outlined earlier and the particular data and model needs in each use case. Evaluate and refine initial pilot studies from the matrix of examples that will likely be enriched by workflow technologies over the next 1-3 years and 4-10 years.
3. Refine initial pilot studies from the matrix of examples that will likely be enriched by workflow technologies over the next 1-3 years and 4-10 years.

Follow-up Processes

1. The workflow research team should carry out a community survey to evaluate the matrix and the pilot design prototype from their point of view.
2. Present results of preliminary design at the AGU/IEEE national meeting(s) to solicit feedback from the larger geoscience computer science communities.

Consensus Processes

1. Appoint and fund an independent community-based evaluation team for assessing and testing the effectiveness of the 3 workflow prototypes.
2. Quantitatively demonstrate how the automated workflow practices have improved the efficiency of the geoscience team (or not) for each workflow prototype
3. Create a national data-software-workflow synthesis center for the geosciences that acts as a clearing house for best practices and further provides post-doctoral and staff support for implementing best practices, hosts visiting researchers and other activities for dissemination of best practices. The synthesis center should support all NSF GEO science communities and not just communities that already have relatively mature data-software-workflow plans in place.

The Workflows Roadmap will include the establishment of a Workflows Synthesis Center for the geosciences that is a national center of excellence and acts as a clearing house for best practices. The Center will support the activities of the Workflows Working Group, by providing post-doctoral and staff support for implementing best practices, hosting visiting researchers, and pursuing community activities for dissemination of best practices. The synthesis center should support all NSF GEO science communities and not just communities that already have relatively mature data-software-workflow plans in place.

Section 5. Status

This section includes a description of the state of the art within the topical area of your roadmap. This should include approaches and technologies from geoscience, cyberinfrastructure, and other fields, the public or commercial sector, etc. that have the potential to benefit the EarthCube enterprise.

In general, it is difficult to assess the current state of the art in the various fields and commercial sectors. This type of assessment needs to happen as part of an ongoing earthcube activity (both because of the scope of the activity and the dynamic nature of the state-of-the art technologies). This activity can be done by the Status and Requirements Task Force through interactions with the science-focused workshops planned for the Fall of 2012.

The Workflows Working Group will include a Status and Requirements Task Force that will: 1) assess the current state-of-the-art in workflow technologies in various domains, 2) evaluate the workflow technology systems available in sciences and the commercial sector and assess their uses, 3) obtain and understand the requirements of the geoscience community, 4) assess the trends and distill major shifts in technologies related to workflows and 5) disseminate their know-how in geoscience workflow needs and state-of-the-art workflow technologies.

5.1 State of the Art in Geosciences

Geoscience applications use an entire gamut of technologies to support their computational pipelines. They range from manual invocation of pipeline steps, through simple scripting solutions, to custom workflow management systems, to general purse workflow management capabilities. There are a number of workflow management systems in use today. Some of them are: DAGMan , Kepler, Pegasus, Wings, and others. In addition to the well-known systems, there are over a hundred of different workflow systems that have been presented in literature. Often times it is not clear which of these systems is used to produce scientific results.

Below we discuss some examples of applications that can benefit from using mainstream workflow technologies.

The Workflows Roadmap will include a Status and Requirements Task Force that will be charged with interacting with the science-focused EarthCube workshops planned for the Fall of 2012..

5.1.2 Current Approaches

At this time, there is no significant penetration of workflow management systems or data cube technology within the national labs like NCAR. For example, in the Research Applications Laboratory at NCAR the majority of research systems are implemented using custom designed infrastructures. Here there are two different paradigms that are in common use.

5.1.2.1 The Glue Script Paradigm

One custom workflow paradigm is to use a scripting language such as Perl, Python, Bourne Shell, etc. to run an individual application module or set of application modules. The underlying scripts used to glue the system together are then typically invoked using a scheduler such as the Unix scheduler cron. This particular paradigm is often used to run numerical models or systems that involve performing numerical model post-processing. The application modules invoked using this paradigm typically perform a well-defined task or set of tasks and then exit. The application modules will then be re-invoked at the next appropriate scheduled time interval. Machines that invoke workflows using the glue script paradigm may often see periods of quiescence in between scheduled workflow run times.

In workflows executed using the glue script paradigm, the custom workflow designers are responsible for making provisions for workflow monitoring, workflow logging, wayward process termination, data file cleanup, resource management, and preventing the invocation of identical tasks.

5.1.2.2 The Control Process Example

In this particular paradigm a control process implemented using a programming language like C++ is executed. The control process in turn is responsible for executing a set of child processes in the workflow. The control process may terminate itself and its children processes after the process family has completed its work. Another common alternative is for the control process to simply monitor the underlying child processes that continuously perform tasks on a data driven or time driven basis.

Workflow designers using the control process paradigm are responsible for making provisions for workflow monitoring, child process re-invocation on failure, workflow logging, wayward child process termination, data file cleanup, resource management and preventing the invocation of identical tasks.

5.2 State of the Art in Cyberinfrastructure

There are a large number of workflow systems available today. We can broadly divide workflow systems into two categories: those that support the composition of standalone application components and those that support the composition of services (Taverna and others [1-8]). Businesses have mostly focused on service-based systems [9]. Expressing applications as services is not applicable in many scientific disciplines where performance and scalability are critical and where the service invocation overheads are simply too expensive. Consequently, scientific workflows are often expressed as workflows of standalone components.

There are many workflow systems that support application component composition and execution. Some of them, such as Triana [10], Kepler [11], VisTrails [12], and XBaya/Airavata [13], provide sophisticated graphical user interfaces for workflow composition However, a major drawback of these tools is that the user is expected to specify the exact locations of files and executables as well as the execution sites. If the state of the cyberinfrastructure changes, for example the target machine goes down, the user needs to modify the workflow. GUIs have limitations because they have trouble capturing large irregular workflows. There are also workflow tools that focus more on the execution of workflows in potentially heterogenous and distributed environments. For example Pegasus, which includes DAGMan, focuses on automatic workflow task scheduling, reliability, and scalability. Pegaus can easily and automatically re-plan and re-schedule in cases of failure because it performs workflow-level checkpointing. When the workflow is restarted, it does not have to redo all the computations from the beginning, just those that have failed. This is particularly useful when workflows are large and compute- and data-intensive. The intermediate data saved during checkpoints can also be used by other workflows.

Some workflow systems differ in terms of component granularity. For example Kepler and Vistrails support code snippets and individual functions and variables, whereas Pegasus workflow components are standalone codes. Furthermore other systems such as Wings also are able to associate rich metadata with data and workflow components. There are also systems such as Swift [14], which take the approach of developing new languages to describe workflow-like computations.

There are also recent efforts to use semantic technologies for workflow composition as can be seen in the Wings system [16]. We are collaborating with Wings to understand the use of ontologies in the specification of workflow components and their data requirements.

5.3 State of the Art in Commercial Sector

There are some examples of use of commercial systems for earth sciences. Some of them include BPEL [3], the Microsoft Workflow Foundation [4], OGC workflow systems [17]. Further investigation in this area is necessary and can be done by the Status and Requirements Task Force.

The Workflows Roadmap will include a Status and Requirements Task Force that will be charged with assessing the current use of commercial systems in geosciences.

References

[1] T. Gunarathne, C. Herath, E. Chinthaka, and S. Marru, "Experience with adapting a WS-BPEL runtime for eScience workflows," presented at the SC-GCE, 2009.

[2] T. Andrews, F. Curbera, H. Dholakia, Y. Goland, J. Klein, F. Leymann, K. Liu, D. Roller, D. Smith, S. Thatte, I. Trickovic, and S. Weerawarana, "Specification: Business Process Execution Language for Web Services Version 1.1," ed, 2003.

[3] "Active BPEL Engine," ed. http://www.active-endpoints.com/active-bpel-engine-overview.htm, 2007.

[4] "Microsoft Workflow Foundation," ed. http://msdn2.microsoft.com/en-us/netframework/aa663328.aspx, 2007.

[5] S. Miles, J. Papay, C. Wroe, P. Lord, C. Goble, L. Moreau, and (2004), "Semantic Description, Publication and Discovery of Workflows in myGrid.," Electronics and Computer Science, University of Southampton. Technical Report ECSTR-IAM04-001, 2004.

[6] A. Slominski, "Adapting BPEL to Scientific Workflows," in *Worfklows for e-Science*, I. Taylor, *et al.*, Eds., ed: Springer, 2006.

[7] T. Oinn, P. Li, D. B. Kell, C. Goble, A. Goderis, M. Greenwood, D. Hull, R. Stevens, D. Turi, and J. Zhao, "Taverna/myGrid: Aligning a Workflow System with the Life Sciences Community," in *Workflows in e-Science*, I. Taylor, *et al.*, Eds., ed: Springer, 2006.

[8] T. Glatard, J. Montagnat, D. Lingrand, and X. Pennec, "Flexible and Efficient Workflow Deployment of Data-Intensive Applications On Grids With MOTEUR," *International Journal of High Performance Computing Applications,* vol. 22, pp. 347-360, 2008.

[9] Microsoft. (2010, *Trident: Scientific Workflow Workbench for Oceanography*. Available: http://www.microsoft.com/mscorp/tc/trident.mspx

[10] D. Churches, G. Gombas, A. Harrison, J. Maassen, C. Robinson, M. Shields, I. Taylor, and I. Wang, "Programming scientific and distributed workflow with Triana services," *Concurrency and Computation: Practice and Experience,* vol. 18, pp. 1021-1037, 2006.

[11] B. Ludascher, I. Altintas, C. Berkley, D. Higgins, E. Jaeger, M. Jones, E. A. Lee, J. Tao, and Y. Zhao, "Scientific workflow management and the Kepler system," *Concurrency and Computation: Practice and Experience,* vol. 18, pp. 1039-1065, 2006.

[12] S. P. Callahan, J. Freire, E. Santos, C. E. Scheidegger, C. T. Silva, and H. T. Vo, "VisTrails: Visualization meets Data Management " in *ACM SIGMOD*, 2006.

[13] Airavata http://incubator.apache.org/airavata/architecture/workflow.html

[14] Y. Zhao, M. Hategan, B. Clifford, I. Foster, G. v. Laszewski, V. Nefedova, I. Raicu, T. Stef-Praun, and M. Wilde, "Swift: Fast, Reliable, Loosely Coupled Parallel Computation," in *2007 IEEE Congress on Services*, 2007, pp. 199-206.

[15] Y. Gil, V. Ratnakar, E. Deelman, G. Mehta, and J. Kim, "Wings for Pegasus: Creating Large-Scale Scientific Applications Using Semantic Representations of Computational Workflows," presented at the Proceedings of the 19th Annual Conference on Innovative Applications of Artificial Intelligence (IAAI), Vancouver, British Columbia, Canada, 2007.

[16] Vijay Kumar, Tahsin Kurc, Varun Ratnakar, Jihie Kim, Gaurang Mehta, Karan Vahi, Yoonju Lee Nelson, P. Sadayappan, Ewa Deelman, Yolanda Gil, Mary Hall, Joel Saltz.Parameterized Specification, Configuration, and Execution of Data-Intensive Scientific Workflows, In Cluster Computing Journal, Vol 13, 2010

[17] OGC Workflows http://www.opengeospatial.org/projects/groups/workflowdwg

Section 6. Solutions

This section describes the process for the identification and comparison (pros and cons) of approaches and technology solutions that will contribute to the EarthCube goal of satisfying current and future research needs of the geoscience end-user.

6.1 The Problem Domain

There are many earth science simulation use cases that appear to leverage different approaches in their inherent workflows. As noted in section 1 (Purpose), discussions regarding this phenomenon of different modes of use of workflows across the user community led to an initial formulation of “Workflow Use Paradigms”. The initial list of identified paradigms were described in Section 1.3.1.

The challenge is that, despite the many inherent workflows that are present in geoscience research areas, a significant proportion of the scientists do not leverage technology mechanisms to automate the discernible workflows in their research work. The underlying issue of lack of workflow automation appears to be more a case of ‘lack of motivation towards technology adoption’ or ‘difficulties in technology adoption’ rather than being an issue of inability to determine the existence of workflows in the geoscientists’ work. This conclusion is discussed in section 3 (challenges) and is supported by the responses to a survey conducted by the workflow community group as detailed below:

In response to the question; “Improving the efficiency of a complicated or error-prone set of computational steps or tasks is a common problem in research. Do you have these types of problems in your research? How do you address them?”, none of the respondents claimed that they *do not* have an ‘error prone set of computational steps or tasks’. Yet their feedback on how they addressed them varied greatly with responses including strategies such as; ignoring, writing down the steps in a notebook or wiki, detailed documentation, writing shell scripts, and creating reusable workflows. Yet only a couple of respondents followed with feedback for the related question of, “what are the strengths and weaknesses of your current approach? What improvements would you like to see?”. Thus, the motivation to adopt technology to solve their problems appears to be low.

These findings indicate that geoscientists need to be guided through the process of discerning the inherent workflows in their research work, and the subsequent automation of those workflows. The survey results also indicate that the approaches towards use of automated workflows in geoscience research would be very diverse and dependent on factors such as,

- the workflow use paradigm
- the current method of dealing with the inherent workflow
- the nature and characteristics of the underlying problem domain
- the stages at which the workflow is apparent in the research process - data collection stage, computational stage, data analysis etc.
- the technical expertise available to the geoscientist or the geoscience research team
- the level of access to cyberinfrastructure

Further, the appropriate approach and suitable technology solutions would be dependent on the underlying objectives of the geoscience research team, the larger geoscience domain or a large organization covering many geoscience areas. Such objectives may include:

- The increased use of existing workflow tools by geoscientists.
- The improvement of existing tools or development of new tools to cover gaps identified by the Status and Requirements Task Force.
- The targeted use of workflow technologies in very high profile projects that aren't currently using them.
- The increased ability to document and reproduce research results and community data products using workflows.
- The extension of workflow technologies or development of new ones to solve grand challenge problems.
- The sustainability of tools.

Many of these factors and objectives are noted in the previous sections of this roadmap document. The processes discussed in this section will be primarily executed by the Engagement Task Force.

The Workflows Working Group will include an Engagement Task Force that will: 1) provide guidance to geoscientists in identifying a approaches to address their workflow needs, 2) assist scientists in evaluating potential workflow technology solutions, 3) request the support of the Status and Requirements Task Force and the Prototyping Task Force when necessary 4) disseminate expertise in workflow solution approaches.

6.2 Processes to Identify Approaches

In many domains beyond geoscience, maturity models have been leveraged as guidance mechanisms to systematically achieve increasing levels of sophistication in desirable capabilities and characteristics. Some domains such as software development adopt a single maturity model such as CMMi across all organizations. Yet some domains such as eGovernment and eHealth tend to specify organization-specific maturity models, borrowing heavily from existing models and customizing them to suit their purposes.

Considering the diverse nature of geosciences and the multifaceted factors listed above, the process of identifying an approach to that would systematically introduce workflow orchestration and automation capabilities in a geoscience research group could be guided by the definition of an organization-specific "Workflow Capability Maturity Model". Such a workflow capability maturity model will define levels of maturity that could would provide the aims and directions of growth in workflow orchestration and automation sophistication within the organization. Each level would detail the milestones and metrics to measure progress. In this manner a workflow capability maturity model will serve as a tool to reflect upon the organization's progress in effectively utilizing workflows to address their problem domain specific orchestration and automation goals.

Support for the creation of such a maturity model for each organization would be provided by Engagement Task Force, to whom the the "task" would be handed off, after the work of the Status and Requirements Task Force.

The Engagement Task Force would be able to help the target organization identify their path of maturing capabilities in workflow orchestration and automation by considering the evident ‘Workflow Use Paradigms’, their unique domain factors and needs,and any specific requirements that may necessitate the involvement of other community groups such as data mining, brokering and interoperability.

6.3 Processes to Identify Technology Solutions

Once the potential involvement of other community groups such as data mining, brokering and interoperability have been established and an appropriate approach to handle the problem domain has been structured through the design of a workflow capability maturity model, the Engagement Task Force will be involved in helping the geoscience team identify technology solutions.

As noted in section 5 (Status), a plethora of wide and varied technology solutions exist, and the process of standardizing and refining these diverse systems, if agreed upon, would be an ongoing and long term process. Thus identifying technology solutions that are aligned with the solution approach to the specific problem domain would be a tricky and non-trivial exercise with the long-term impact. In addition to working with the other community groups that would be associated in this effort, the requisite standards that should be adopted will also need to be determined.

Therefore, the process to identify technology solutions should be done in a systematic manner through the definition of a “Technology Evaluation Framework”. Such a framework should be designed for each technology solution related goals or tasks that evolve from the progress through the workflow capability maturity model. A technology evaluation framework would specify the comparison characteristics and associated metrics, including absolute conditions that need to be satisfied by a particular technology solution as well as specifics with respect to the process of evaluation such as requisite testbeds, infrastructure, prototyping needs etc.

The Engagement Task Force will provide the necessary guidance for this, and when necessary, will obtain the support and assistance of other task forces. For example, the Prototyping Task force would most likely be needed in the cases of multiple technology solution evaluation or in the cases of complex technology solution integration.

There will be a need to assess the adoption and impact of workflow technologies as a result of EarthCube activities. This will require defining metrics to measure adoption and impact, and to collect data as EarthCube progresses. In the early stages of the project, baseline data should be collected to assess the state of the art and the initial levels of adoption. As the activities of the Workflows Working Group progress, additional data should be collected to assess adoption and impact. The data should be analyzed, particularly to detect barriers and challenges that will inform and adjust the roadmap activities.

The Assessment Task Force will track and assess the impact of workflow technologies across geosciences through: 1) defining metrics to measure impact in geosciences, 2) collecting quantitative and qualitative data at the early stages of EarthCube as baselines, 3) collecting additional data to measure and demonstrate progress as the roadmap activities progress, 4) analyzing the collected data to understand any issues and challenges that may need to be resolved in order to achieve the roadmap goals.

Section 7. Process

This section describes the process(es) to develop community standards, protocols, test data, use cases, etc. that are necessary to mature the functionality of the topical area and promote interoperability and integration between elements of EarthCube.

This section relies on previous sections and also on Section 9, Management.

7.1 Overview

Key elements of the processes to be undertaken include:

- Defining workflow types and capabilities, drawn from community requirements.
- Identifying relevant supplemental standards from existing standards-making bodies (OGC, ISO, W3C). These can include data formatting standards, transfer protocols, services and their APIs, metadata encoding standards, etc.
- Classifying software useful for workflows. This should include capabilities and scope and also take into consideration project management and governance, such as OSSWatch evaluations.
- Identifying gaps in standards, workflow tool capabilities, and underlying infrastructure.
- Supporting the use of workflow technologies through training, documentation, outreach, etc.
- Directly supporting NSF (and other) research organizations through advanced support requests that are matched with the appropriate workflow experts.
- Facilitating new collaborations.
- Building scientific and technical bridges to other areas (semantics, data management, etc).

The EarthCube Workflow Working Group will be charged with developing and executing these processes. The Workflow Working Group must adopt an appropriate governance model (Section 9).

7.2 Processes to Develop Community Standards

There are already standard making organizations (OGC, ISO, W3C, NIST), and defacto standards (NetCDF, OPeNDAP). The EarthCube Workflow Working Group will not serve as a new standards-defining body. It will instead provide guidance to the community on existing standards relevant to workflows and underlying services, and it may also provide advice for participating in existing standard-making organizations. It will also monitor standards-making organizations who are developing workflow-related standards and participate.

The Workflow Roadmap will include the creation of a Community Interaction Task Force that will be responsible for documenting existing standards and making recommendations on appropriate standards to specific research groups.

7.3 Processes to Develop Community Use Cases

A recurring problem identified in our workshops was the low adoption rate of workflow technologies even though the potential value was understood. This could result from several causes: the technologies are not known, they are perceived to be too cumbersome or complicated, the researchers are too busy to learn how to use them, the technologies are too general and need to be populated with more geoscience-centric examples, etc.

The Workflow Roadmap will include the creation of a Community Interaction Task Force that will be responsible for documenting use cases for current and potential uses of workflows in geosciences.

7.4 Processes to Develop Community Reference Implementations

An important problem in the cyberinfrastructure community has been the lack of community-developed reference implementations for many standards. This has the potential to produce undesirable results such as standards controlled by a single implementing group or a fragmented development community attempting to implement the same standards independently.

The Workflows Community Group recommends the adoption of Apache Software Foundation-style open community processes (preferably within the Apache organization) for developing reference implementations of community standards. The Apache way has the advantage of creating a merit-driven, neutral development area that is, in addition to being open source, required to demonstrate developer diversity.

7.5 Promoting Interoperability and Integration

Standards assist with interoperability and integration but must be supplemented by requirements, driving use cases, community reference implementations, testbeds, and plans for transfer to operations. See Section 4, “Requirements”, for more information on requirements. A number of existing organizations, such as the OGC, already have considerable experience in these area, so the Workflows Community should partner with these groups. EarthCube-wide integration efforts will also need to occur. We expect to establish or participate in integration efforts with other EarthCube activities, particularly the Data Discovery, Mining, and Integration community, the Semantics community, the Brokering concept award, the Web Services concept award, and the Interop concept award. See Section 2, “Communications”, and Section 9, “Management” for more information on liaisons with these other EarthCube entities.

Section 8. Timeline

This section describes the timeline for the project and all related sub-projects, including prioritization of activities and measurable milestones/major achievements and total resources (human and financial) required to achieve roadmap goals over a period of the next 3 to 5 years.

The overarching goal of the workflows working group is to make workflows ubiquitous within the geosciences community. This roadmap is motivated towards addressing the challenges we see in the community as extensively discussed in the previous sections.

The biggest issues against achieving the overarching goal of ubiquity is the lack of awareness on how to map science challenges into workflow technologies that would improve the process, and the diverse and dispersed workflow community.

Hence our activity prioritization, our milestones and the associated timelines are heavily influenced towards addressing the major issues early on, in the roadmap execution.

8.1 Prioritization of Activities

Once the organizational structure has been established, each structural component will be tasked with a set of prioritized activities. In order to ensure appropriate focus and consistent action towards achieving the workflow working group's short term, medium term and long term goals, we have categorized 3 thrust areas for activities. These thrusts are, Awareness, Enabling and Infrastructure & Services. Many important activities within these thrusts have been listed below. However, it should be noted that additional activities would be generated after the workflow working group's foundational structural components and substructural components are established and begin operations.

In terms of prioritization, many, though not all of the activities within the Awareness Thrust will have the highest priority. Next in priority would be many of the activities within the Infrastructure and Services Thrust. In order to assign specific priorities across the Thrust areas, each activity listed below is followed by an alphabetical character denoting its relative priority, with "A" being the highest priority, "B" being the next in priority etc.

Awareness Thrust

The main goal will be on projecting what workflows are, learning more about their current and potential uses. The associated activities include:

- Create a compendium of "workflow paradigms" **- A**
- Create a compendium of current geo-workflow success stories **- A**
- Create a compendium of identified opportunities for adoption **- B**
- Create awareness-related tutorial materials **- A**
- Actions to form and develop the workflows community in the short-term:
    - Reach out to people already using workflows and identify their issues **- A**

    - Reach out to geoscientists in major geoscience research centers and identify why they do not use workflows **- A**
    - Review findings from outreach initiatives a formulate an action plan that would be executed by relevant task forces **- B**

Enabling Thrust

- Services for implementing earth science workflows in the medium-term:
    - Select opportunities for the technology
        - Create a community approaching center **- C**
        - Create a center pursuing strategic opportunity areas **- D**
    - Perform collaborative prototyping **- C**

Infrastructure and Services Thrust

- Support workflow community: mailing lists, discussion boards, etc. **- A**
- Create a catalog of open source workflow systems and tools **- B**
- Create and maintain a catalog of workflows for many geo disciplines **- B**
- Support open sharing models associated data preparation workflows **- C**
- Identify reusable workflow services: find, reuse, adapt **- D**
- Identify capabilities for execution through **- B**
    - NCAR, NCEP resources, etc
    - Campus Clusters
    - national resources like XSEDE, OSG, FutureGrid, etc
- Implement data ingest plugin to pull in data from national data providers **- B**
    - USGS, MODIS and such
- Workflow publication and citation services **- B**
- Metadata and provenance publication **- C**
- Provide server side - client side visualization support **- C**

8.2 Measurable Milestones

Many of the measurable milestones will emanate from the mandates of the task forces, once they are established and create their operational plans. The following is an initial list of measurable milestones:

- Areas in geosciences that have active participation in the workflow group activities
- Number of awareness tutorials conducted in face to face meetings
- Number of attendees to on-line tutorials about wokflows
- Number of workflow vignettes collected from the community
- Number of active members in the workflow community who have contributed to online discussions
- Number of research challenges documented in group reports that drive research in workflows
- Number of specific science questions where using workflows made a difference to the scientists
- Time saving estimates due to workflow reuse
- Number of times that models are reused in different workflows

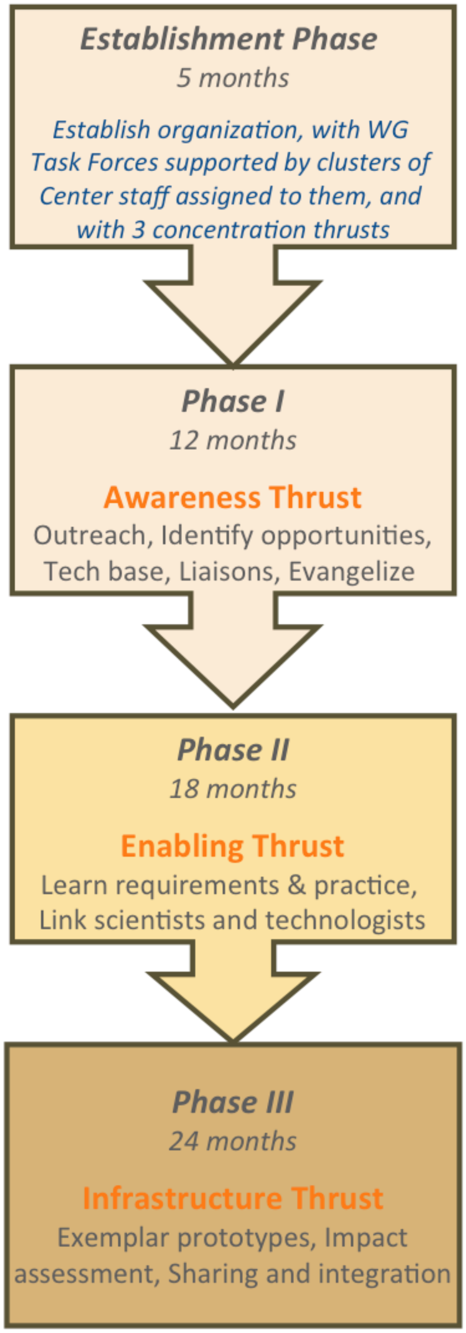
- Number of documented success stories of workflows in geosciences resulting from the group's activities
- Number of cross-disciplinary analyses carried out with workflows
- Number of projects that di analyses that are larger scale, faster, or better through workflows
- Number of projects that include workflow provenance records linked to their publications
- Number of data sources accessed from workflows

8.3 Timeline of Projects and Subprojects

The timeline for establishing the foundational structures in this roadmap is listed below. It is assumed that this timeline begins on the date that the funded Workflows Working Group would be established within the EarthCube initiative. It is also assumed that once established, these foundational structures will remain operational for the next 3-5 years unless retired by the Steering Committe.

- Months 1-5: Establishment Phase
    - Steering Committee and basic on-line operational infrastructure: 1 month
    - Task Forces: established within 3 months
    - Synthesis Center: Initial operations within 5 months
- Months 6-18: Awareness Phase: The awareness thrust will be emphasized
- Months 18-36: Enabling Phase: The enabling thrust will be emphasized
- Months 37-60: Infrastructure and Services Phase: the infrastructure and services thrust will be emphasized

A diagram of the timeline is shown below.


Establishment Phase
5 months
Establish organization, with WG Task Forces supported by clusters of Center staff assigned to them, and with 3 concentration thrusts
Phase I
12 months
Awareness Thrust
Outreach, Identify opportunities, Tech base, Liaisons, Evangelize
Phase II
18 months
Enabling Thrust
Learn requirements & practice, Link scientists and technologists
Phase III
24 months
Infrastructure Thrust
Exemplar prototypes, Impact assessment, Sharing and integration

8.4 Major Achievements

The following would be considered to be significant achievements of the Workflow Working Group.

- Establishing a diverse and cross-disciplinary Steering Committee that worked with the community to establish a viable path to accomplishing the goals of the roadmap
- Establishing and operationalizing the Workflow Synthesis Center with measurable impact on adoption of workflows and best practices across geosciences

- Establishing and operationalizing the Workflow Group Task Forces with strong participation from the community
- Creating an online workflows community with active member participation and tight integration with other EarthCube and relevant community activities

8.5 Total Resources Required

The Workflows Synthesis Center will require resources to support Steering Committee members, several full-time personnel (post-docs and research programmers), a visitor program, and travel for participating in community activities.

Section 9. Management

This section describes the management / governance / coordination plan and decision-making processes necessary to successfully establish standing committee(s) and subcommittees (if warranted), including a plan to identify and respond to shifts in technologies and changing needs at the end-point of use. It includes discussion of approaches to educating end-users and achieving community consensus on advancing the capability/technological solution.

9.1 Management, Governance, and Coordination Goals

The goal of management is to execute the roadmap and its principal goal of making workflows ubiquitous within the geosciences.

The working group management must be made as efficient and effective as possible. Traditional management processes are inadequate for the roadmap evolution and execution, since we must coordinate a confederacy of independent organizations and individuals. Thus the problems that we need to solve can be categorized into two perspectives - organizational and individual.

The challenges that need to be addressed from an organizational perspective include,

- Establishing an effective organizational structure that would enable efficient strategizing;
- Establishing an effective operational structure that would ensure smooth and timely operational activities;
- Establishing effective processes for creating groups, organizations, etc.;
- Establishing effective processes to facilitate consensus and enable efficient decision-making.

The problems that need to be addressed from an individual's perspective include,

- Efficient and productive use of participants' time;
- Creating incentives beyond funding to encourage participation;
- Supporting and rewarding initiative by individuals;

As discussed below, the organizational goals will primarily be achieved through the organizational structure of the Workflows Working Group and the associated open community process model. The goals from an individual's perspective will to a large extent, be facilitated by the substructures within the overall organizational structure, and the associated open community process model.

9.2 Management, Governance, and Coordination Plan

The strategy will be to establish a central Steering Committee for the Workflows Working Group with the flexibility that allows its members to take initiative to address problems. The group would include specific Task Forces that would address the challenges and enable the implementation of the strategies proposed in this roadmap. The roadmap also envisions establishing an institute that would function as a Synthesis Center. Each of these structural components and their operational processes and primary responsibilities will be discussed in detail below.

The Steering Committee will be empowered, through votes by its members, to create additional organization structures as deemed necessary to carry out the working group's mission.

The Workflow Working Group community will operate on an open community model. The processes that govern this model are detailed below. They are influenced by the highly successful mechanisms within the Apache Software Foundation's open community model (http://www.apache.org/foundation/how-it-works.html).

Members of the Workflow Working Group will initially include funded participants, but it will establish a mechanism on the acceptance of new members, such as being voted in by the Steering Committee. Criteria for nominations of new members will be left open but guidelines will be provided.

The Workflow Working Group will leverage online tools and services to facilitate and manage its mission. The tools will include;

- An open mailing list that will serve as the primary communication mechanism for the working group. More specific mailing lists may be created if deemed necessary;
- A web site editable by members of the Steering Committee that will provide access to all working group documents;
- A web-based document repository (with version tracking) with open read access and write access by request; and
- An issue tracking system that with open write access that will support the creation, assignment, progress, and resolution of hierarchically defined issues.

Thus the information generated within and by these tools will be by default, publicly readable, with write access granted by Steering Committee determination.

In addition to the plans described above, the recommendation of the EarthCube Governance Working Group about appropriate organization guidelines will be given due consideration and blended in with our proposed plan.

9.3 Decision-Making Processes

All decisions will be made by public vote on an open Steering Committee mailing list. Each Steering Committee member will have one vote. Votes will be open for a specific period (typically 72 hours) and will have an accompanying discussion period. Members will allowed to express "no opinion", or "+/-0" votes, which are counted toward the quorum.

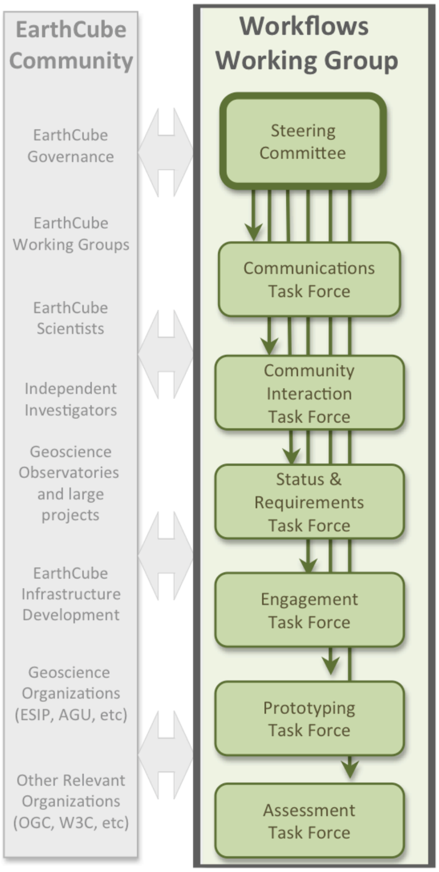
Measures will be deemed to pass if, 1) a quorum of Steering Committee members has voted within the set period, and 2) the number of positive votes exceeds the number of negative votes. All decisions will be conveyed through open, archived discussions on mailing lists.

9.4 Establishing Organization Substructures

The Workflows Working Group will be the driver of EarthCube workflow activities as well as interactions with the EarthCube community. The roadmap recommends the creation of a Workflows Synthesis Center that will support the Workflows Working Group. In this section, we describe the organizational structure of each of them as well as their interrelated functions.

9.4.1 Organization of the Workflows Working Group

The Workflows Working Group will be structured with several Task Forces that will interact with the EarthCube community. The roadmap defines six Task Forces with clearly distinguished functions as well as the interactions among them. The Task Forces are summarized in the diagram below.


EarthCube Community
EarthCube Governance
EarthCube Working Groups
EarthCube Scientists
Independent Investigators
Geoscience Observatories and large projects
EarthCube Infrastructure Development
Geoscience Organizations (ESIP, AGU, etc)
Other Relevant Organizations (OGC, W3C, etc)
Workflows Working Group
Steering Committee
Communications Task Force
Community Interaction Task Force
Status & Requirements Task Force
Engagement Task Force
Prototyping Task Force
Assessment Task Force

The Task Forces and their responsibilities are described in detail elsewhere in this roadmap document, and are summarized next.

Communications Task Force
Communications Task Force will disseminate information to the community through: 1) a web presence with accessible (non-technical) information for geosciences researchers, 2) clear points of entry and participation for engaging newcomers, 3) pursue synergistic and opportunistic meetings of for both virtual and physical participation, and 4) measure the growth of the participating community and their engagement over time.

The Communications Task Force will organize and actively participate in community-wide EarthCube workshops and other significant community events by: 1) defining goals and expected products from a workflows session at the workshop, 2) preparing materials to facilitate engagement from workshop participants, 3) establishing strategic lines of communication with that community that are sustainable in the longer term 4) aligning with events involving cyberinfrastructure development such as AGU, XSEDE annual conferences and Supercomputing.

The Communications Task Force will pursue publications through: 1) submitting regular reports in newsletters and journals regarding group activities and progress, 2) encouraging group participants and the community to publish significant accomplishments, 3) collecting published materials and making them publicly available.

Community Interaction Task Force
The Interaction Task Force will pursue and coordinate interactions with major funded projects, standards efforts, government agencies, and community groups in geosciences to: 1) uncover use cases and opportunities for dissemination of workflow technologies, 2) collect success stories, lessons learned, and future requirements and challenges for workflows in geosciences, 3) leverage their efforts by aligning them with the Workflows roadmap, and 4) take advantage of opportunities to participate in events where large communities already participate

The Interaction Task Force will engage computer science researchers with EarthCube activities will benefit the geosciences community by enabling: 1) faster transfer of new advances in basic workflow research into geoscience infrastructure, and 2) defining basic research needed in workflows motivated by grand challenges in geosciences.

Status and Requirements Task Force
The Workflows Working Group will include a Status and Requirements Task Force that will: 1) assess the current state-of-the-art in workflow technologies in various domains, 2) evaluate the workflow technology systems available in the commercial sector and assess their uses, 3) obtain and understand the requirements of the geoscience community, 4) investigate and compare the workflow technologies that significantly overlap in the workflow requirements they address, 5) assess the trends and distill major shifts in technologies related to workflows and, 6) disseminate their know-how in geoscience workflow needs and state-of-the-art workflow technologies.

Engagement Task Force
Engagement Task Force will; 1) provide guidance to geoscientists in identifying a approaches to address their workflow needs, 2) assist scientists in evaluating potential workflow technology solutions, 3) request the support of the Status and Requirements Task Force and the Prototyping Task Force when necessary, and 4) disseminate their expertise in workflow solution approaches.

Prototyping Task Force
The Prototyping Task Force will pursue prototypes of workflow technologies in various domains of geosciences can: 1) provide exemplary demonstrations of the capabilities of workflow technologies for doing science, 2) gain advocates among scientists that benefit from workflow technologies, 3) uncover difficulties in adoption and need for outreach activities, 4) disseminate expertise to manage this novel aspect of cyberinfrastructure, 5) create opportunities for interactive educational materials for students in their geosciences domain of interest, and 6) produce scientific advances and discoveries that would not occur as fast or as easily without workflows.

Assessment Task Force
The Assessment Task Force will track and assess the impact of workflow technologies across geosciences through: 1) defining metrics to measure impact in geosciences, 2) collecting quantitative and qualitative data at the early stages of EarthCube as baselines, 3) collecting additional data to measure and demonstrate progress as the roadmap activities progress, 4) analyzing the collected data to understand any issues and challenges that may need to be resolved in order to achieve the roadmap goals.

9.4.2 Organization of the Workflows Synthesis Center

A Workflows Synthesis Center will be established as a national center of excellence to efficiently and effectively facilitate the multitude of activities and operational tasks discussed in this roadmap.

The center will leverage upon existing organizations in order to create a distributed center which is strongly connected with tight communication mechanisms.

The center will have 2-3 co-directors that will manage the activities of the center and will coordinate with the Steering Committee of the Workflows Working Group.

The center will comprise postdoctoral students and full time programming staff. It could be initially created with 5 FTEs spanning both technical and science expertise.

The personnel in the center will be organized into six clusters that will support the activities of the six corresponding Task Forces of the Workflows Working Group. The clusters will take guidance and priorities from the Task Forces.

The center will host visitors from research groups across geosciences. These visitors will work with center personnel through training and prototyping activities.

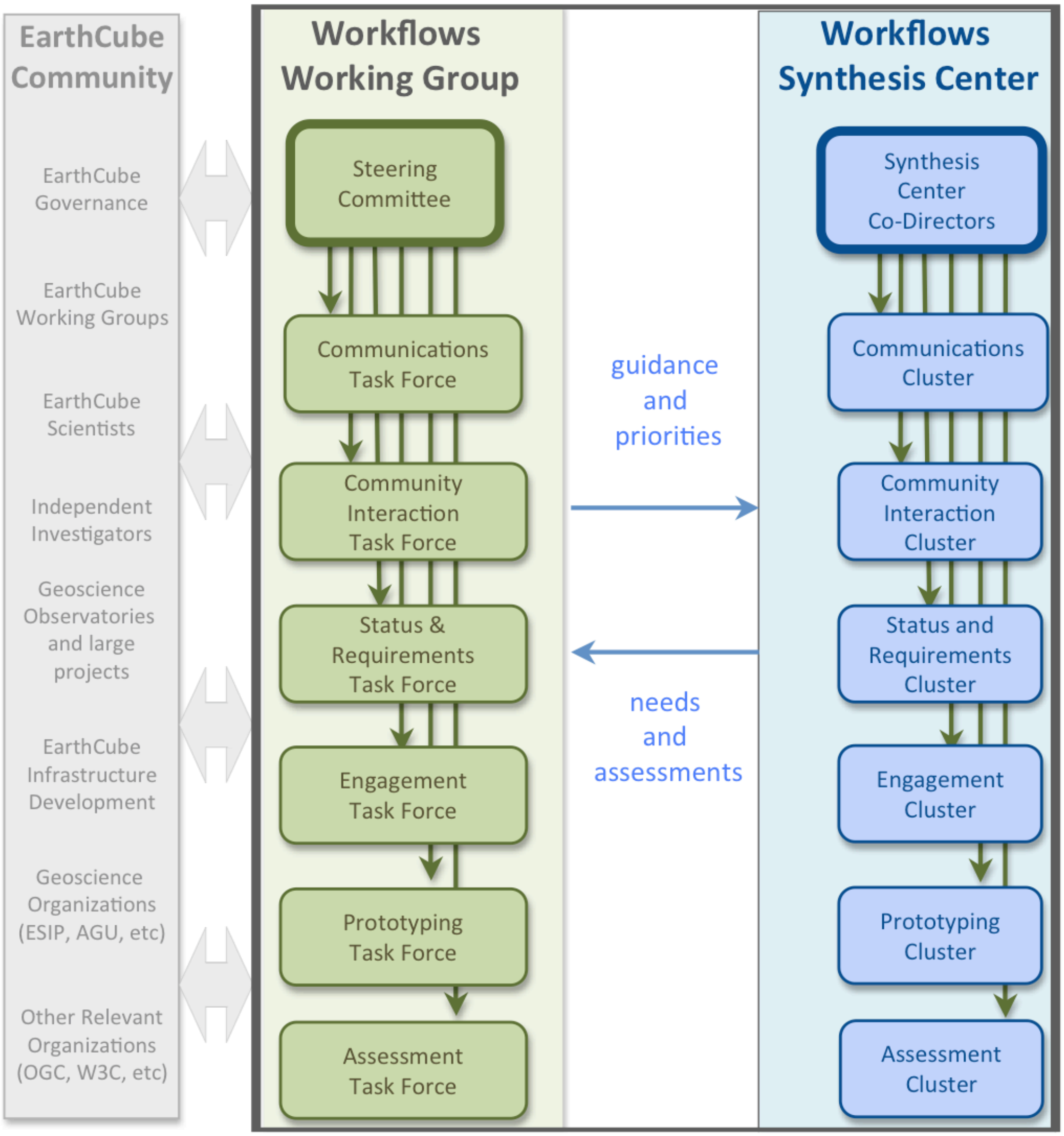
There are many tasks that the Workflows Synthesis Center will facilitate or that its personnel will be involved in, arising from the mandates of the Task Forces of the Workflows Working Group. These tasks include:

- developing introductory material, tutorials, and use cases
- attending geoscience meetings and evangelizing the workflow community mission
- hosting visitors from geoscience groups to helping them prototype workflow applications
- visiting project sites to engage with geoscientists in their labs/centers/field-sites for 1-2 weeks at a time
- organizing and conducting summer/winter schools
- creating and maintaining a knowledge base of relevant FAQs
- collecting and tracking use cases/success stories and feedback from end users
- organizing outreach workshops
- developing approaches to “workflow citation” (analogous to data citation), linking workflows to publications, etc.
- recommending the creation of new task forces or the merge of task forces to the steering committee
- coordinating and participating in ongoing activities of other groups or organizations such as IEDA (Integrated Earth Data Applications)
- reflecting on prior initiatives/activities of the workflow working group task forces and analyzing their impact on workflows
- coordinating with government organizations and other organizations that have geodata centers
- coordinating with other EarthCube working groups

The Workflows Roadmap will include a Workflows Synthesis Center as a national center of excellence that will be geographically distributed and will: 1) include senior researchers, post-doctoral students, and research programmers with expertise in workflow technologies and some geosciences background; 2) conduct courses for students about workflow research and for developers about workflow software adoption; 3) host visitors who will engage in training and prototyping activities with center personnel.

9.4.3 Overall Organizational Structure

An overview of the envisioned organization structure is illustrated in the following diagram. The Workflows Working Group will coordinate interactions with the broader EarthCube community, and draw its members from it. It will provide guidance on requirements and priorities to the Workflows Synthesis Center. In turn, the Workflows Synthesis Center will support the activities and Task Forces of the Workflows Working Group, while providing information about the needs of the Center to support those activities as well as assessments of the viability of the Workflows Working Group activities. The Workflows Working Group will have six Task Forces with corresponding clusters in the Workflows Synthesis Center.


EarthCube Community
EarthCube Governance
EarthCube Working Groups
EarthCube Scientists
Independent Investigators
Geoscience Observatories and large projects
EarthCube Infrastructure Development
Geoscience Organizations (ESIP, AGU, etc)
Other Relevant Organizations (OGC, W3C, etc)
Workflows Working Group
Steering Committee
Communications Task Force
Community Interaction Task Force
Status & Requirements Task Force
Engagement Task Force
Prototyping Task Force
Assessment Task Force
guidance and priorities
needs and assessments
Workflows Synthesis Center
Synthesis Center Co-Directors
Communications Cluster
Community Interaction Cluster
Status and Requirements Cluster
Engagement Cluster
Prototyping Cluster
Assessment Cluster

9.5 Responding to Shifts in Technologies

The workflows working group includes a “Status and Requirements Task Force” as described above. Their mandate includes assessing the trends and major shifts in technologies related to workflows. Based on their recommendations, the Steering Committee will have the authority to amend the roadmap and its derived activities, or establish task force to evaluate the impact of major shifts in technologies.

9.6 Responding to Changing Needs

The workflows working group includes a “Status and Requirements Task Force” and an “Engagement Task Force” as described above. Their mandates include, obtaining and understanding the requirements of the geoscience community and providing guidance to geoscientists in identifying a approaches to address their workflow needs. If these task forces note major changes in the workflow needs of geoscientists, based on their recommendations, the Steering Committee will have the authority to amend the roadmap and its derived activities, or establish task force to formulate responses to the changing needs.

9.7 Approaches to Educating End Users

The Steering Committee will engage with end users through processes described in Section 2 (Communications). Additionally, the workflows community group includes two task forces, the Communications Task Force and the Community Outreach Task Force, whose mandates include approaches to educating end users.

9.8 Approaches to Achieving Community Consensus on Advancing Workflows

The Steering Committee will be composed of individuals representing a broad range of stakeholders, including federal agencies, representatives from NSF-funded groups, and individual scientists. When necessary, as discussed above, consensus will be achieved by Steering Committee votes.

Section 10. Risks

This section identifies risks and additional challenges to the successful establishment of any working group on workflows, and any unique risks associated with a working group on workflows. It also describes approaches to mitigate identified risks.

10.1 Potential Risks in Establishing a Working Group on Workflows

10.1.1 Not Establishing Meaningful Requirements

One of the initial risks in establishing a working group on workflows is to fail to establish meaningful requirements. The scientific user community utilizes workflow technology either explicitly or implicitly. In the majority of cases, the workflow technology used by the scientific community is custom developed by scientists or associated staff. This technology will often meet the minimum needs of the scientists but may lack attributes that would make the technical solution more robust and useful. In order to make workflow technology useful, it is essential to understand and ascertain the underlying needs of the user community.

10.1.2 Substantive Differences in User Requirements

Even if meaningful requirements are elucidated, there will be substantive differences in user requirements. Operational users will likely have substantially different needs from researchers, since the former may care more about robustness and scale while the later may focus more on exploration and flexibility. These differences may also manifest across different research areas, paleontologists may have substantially different needs from numerical modelers in meteorology. Identifying appropriate workflow technologies will be key to success.

10.1.3 Not Addressing Workflow Requirements

Once scientific workflow requirements are well-understood, the next challenge/risk for the workflows group will be how to address the stated requirements. Here the response or action plan developed by the workflows group may prove inadequate. A related risk is not having sufficient staff to address the requirements.

A related risk is a wrong prioritization of the requirements. The roadmap will result in more requirements than can be addressed, so priorities will need to be set. This may not be done optimally.

10.1.4 Inadequate Communication with the Scientific User Community

If there is inadequate communication between the workflow requirements group and the scientific user community, then it is unlikely that meaningful requirements will be established and that workflow technology will be adopted. The lack of communication can come from either the side of the workflows working group or the side of the user community. For example, the workflows group may not be aware of or may not gain access to data/models in the geosciences community. On the other hand, the geosciences community may not participate sufficiently in EarthCube research and development.

10.1.5 Lack of Adoption

Even if meaningful requirements are established and adequate channels of communication are established with the user community, workflow technology may still run into adoption difficulties. New technology adoption may be costly in time and resources; requires a change in work habits and a change in work culture; and requires advocacy by staff and management within an organization. Projects, groups and organizations must first see an advantage in using the new technology. They then have to identify individuals interested in adoption and must be willing to allocate time to perform a test and evaluation of the new technology within a project of sufficiently limited scope to address schedule risk. Individuals who have performed the trial adoption must be willing to make a prudent assessment of the software and the assessment of the software must be sufficiently positive. If the workflow technology is difficult to use or otherwise is found lacking in promise, the assessment will be negative.

Finally, even if the assessment is positive, there must be sufficient advocacy within the organization for broad usage.

10.1.6 Choosing Wrong Software Engineering Methodology

Software development models need to be fit to the project. For example, waterfall-based model are implicit in many project organizations but may not match well with EarthCube's scope, diversity, and timeline. Agile approaches (with frequent iterations and smaller project units) may be better suited. Failure to identify and implement the correct methodology at the beginning of the project risks inefficient use of resources, producing solutions that are not a good match to time-evolving requirements.

10.2 Risk Mitigation Strategies

10.2.1 Creating an EarthCube Workflow Working Group

Our overarching risk mitigation strategy is to create an EarthCube Workflow Working Group with an appropriate governance model (see Section 9). The success of this working group will depend upon

- Attracting a large and diverse set of participants from the geosciences. Most of these participants will be volunteering their efforts, so proper incentives will need to be established.
- Establishing mechanisms within this group for decision making such as prioritizing efforts and resources.

10.2.2 Creating a Widespread Workflow Technology User Community

In order to generate a critical mass of workflow technology users, it would be helpful to identify different sets of significant users who would help to usher the technology to the general community. Trial projects for utilizing the workflow technology in conjunction with Earth Cube technology could be funded to establish footholds in the user community.

10.2.3 Developing Channels of Communication Between the User Community and the Working Group

It will be important to encourage channels of communication between the user community and the Workflows working group. Online forums,support desks, conference forums, and the like will all be helpful in promoting workflow technology use. See Section 2, “Communications”, for the Workflows Roadmap communication plan.

10.2.4 Promoting Ease of Use

It will be important to ensure that the workflow technology being promoted is easy to install and easy to use. The technology must also be stable and reliable. If the technology is difficult to install, difficult to use or is not stable/reliable, adoption will be difficult. In order to promote ease of use, there should be adequate documentation, tutorials and example workflows to help potential users get started using the technology.

10.2.5 Acquiring User Feedback

Mechanisms such as online forms or user questionnaires should be employed to assist in acquiring user feedback. It will be helpful to organize the feedback into categories in order to get a better handle on the strengths and weaknesses of the current workflow technology.

10.2.6 Responding to User Feedback

When getting user feedback, it will be important for workflow technology developers to effectively respond to the feedback.

10.2.7 Providing Effective User Support

It will be important for workflow technology providers to provide effective user support and follow up. Here a set of best practices would be helpful.