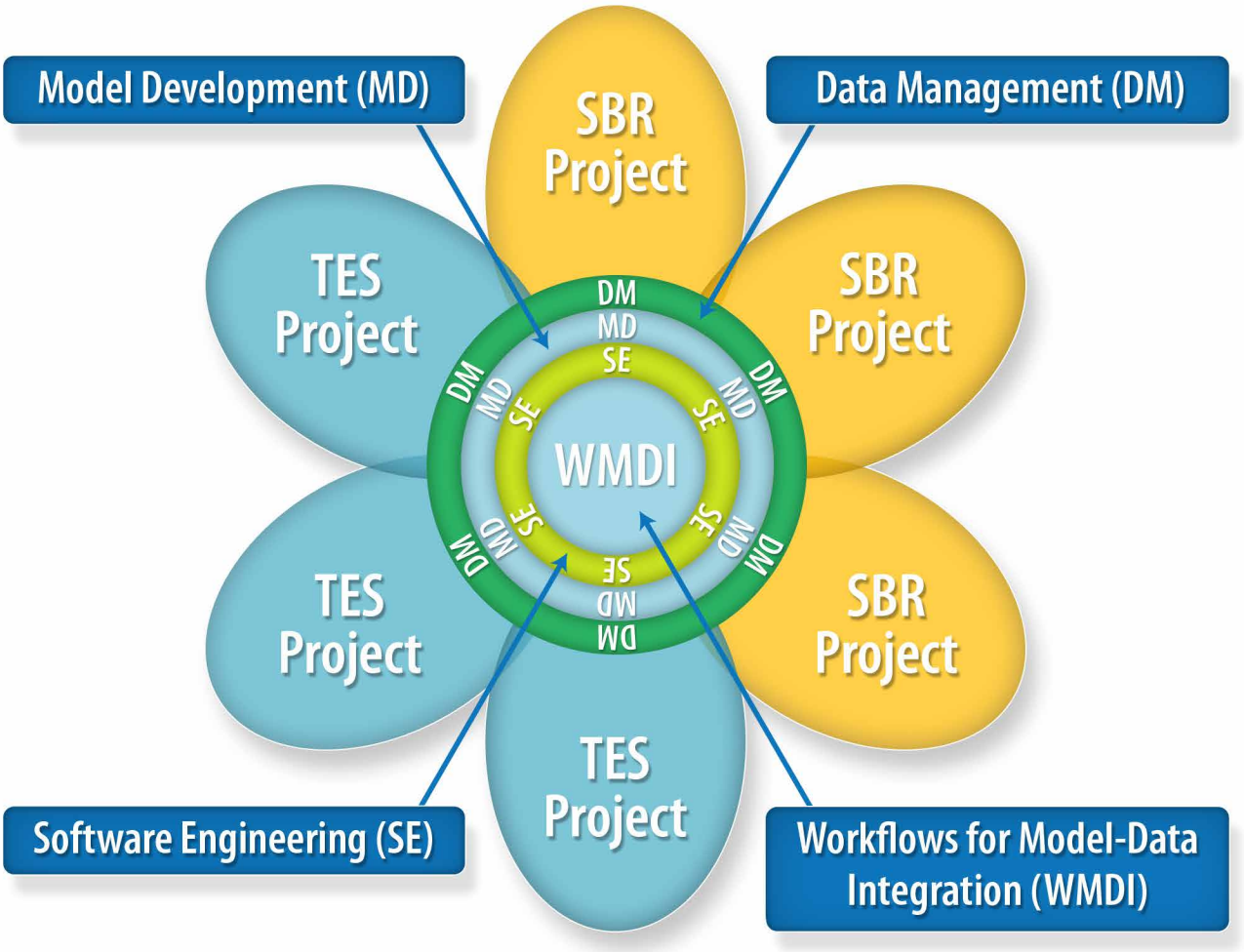
DOE/SC-0178

Building a Cyberinfrastructure for Environmental System Science: Modeling Frameworks, Data Management, and Scientific Workflows

Workshop Report



Environmental System Science

U.S. DEPARTMENT OF ENERGY | Office of Science

Office of Biological and Environmental Research

Environmental System Science Workshop on Model-Data Integration: Modeling Frameworks, Data Management, and Scientific Workflows

April 30–May 1, 2015
Potomac, Maryland

Convened by

U.S. Department of Energy Office of Science
Office of Biological and Environmental Research

ORGANIZING COMMITTEE

Workshop Chair: David Moulton
Los Alamos National Laboratory

Software Engineering

David Moulton
Los Alamos National Laboratory

Dean Williams
Lawrence Livermore National Laboratory

Data Management

Deb Agarwal
Lawrence Berkeley National Laboratory

Tom Boden
Oak Ridge National Laboratory

Roelof Versteeg
Subsurface Insights, Inc.

Model Development

Charlie Koven
Lawrence Berkeley National Laboratory

Tim Scheibe
Pacific Northwest National Laboratory

Carl Steefel
Lawrence Berkeley National Laboratory

Peter Thornton
Oak Ridge National Laboratory

WRITING TEAM

David Moulton
Los Alamos National Laboratory

Tim Scheibe
Pacific Northwest National Laboratory

Haruko Wainwright
Lawrence Berkeley National Laboratory

Roelof Versteeg
Subsurface Insights, Inc.

Carl Steefel
Lawrence Berkeley National Laboratory

Peter Thornton
Oak Ridge National Laboratory

Scott Painter
Oak Ridge National Laboratory

Deb Agarwal
Lawrence Berkeley National Laboratory

ORGANIZERS

Climate and Environmental Sciences Division

David Lesmes
David.Lesmes@science.doe.gov

Justin Hnilo
Justin.Hnilo@science.doe.gov

Mission: The Office of Biological and Environmental Research (BER) advances world-class fundamental research programs and scientific user facilities to support the Department of Energy's energy, environment, and basic research missions. Addressing diverse and critical global challenges, the BER program seeks to understand how genomic information is translated to functional capabilities, enabling more confident redesign of microbes and plants for sustainable biofuel production, improved carbon storage, or contaminant bioremediation. BER research advances understanding of the roles of Earth's biogeochemical systems (the atmosphere, land, oceans, sea ice, and subsurface) in determining climate so that it can be predicted decades or centuries into the future, information needed to plan for energy and resource needs. Solutions to these challenges are driven by a foundation of scientific knowledge and inquiry in atmospheric chemistry and physics, ecology, biology, and biogeochemistry.

About the cover: Envisioned is a community-driven cyberinfrastructure supporting BER Environmental System Science (ESS) activities to facilitate the iterative cycle of model-driven experimentation and observation and accelerate scientific discovery. This cyberinfrastructure will be developed in phases by a dynamic and coordinated set of working groups with expertise in model development, data management, software engineering, and workflows for model-data integration. Working groups will be formed by an overarching executive committee as high-priority needs are identified and will be dissolved when the specific tasks are completed. The executive committee will consist of representatives from major projects funded by the Terrestrial Ecosystem Science (TES) and Subsurface Biogeochemical Research (SBR) programs, which constitute BER's ESS activity. The committee will help to identify and prioritize the topics that the working groups address. Pursuit of these topics by the working groups will be based on specific use cases selected from existing ESS projects and designed to be of general utility to the broader ESS community.

Suggested citation for this report: U.S. DOE. 2015. *Building a Cyberinfrastructure for Environmental System Science: Modeling Frameworks, Data Management, and Scientific Workflows; Workshop Report*, DOE/SC-0178. U.S. Department of Energy Office of Science (doesbr.org/ESS-WorkingGroups/).

DOE/SC-0178

Building a Cyberinfrastructure for Environmental System Science: Modeling Frameworks, Data Management, and Scientific Workflows

Workshop Report

Published: November 2015

U.S. DEPARTMENT OF ENERGY | Office of Science

Office of Biological and Environmental Research

Table of Contents

Executive Summary

The Environmental System Science (ESS) activity within the Department of Energy's (DOE) Office of Biological and Environmental Research (BER) seeks to advance a robust, predictive understanding of terrestrial environments, extending from bedrock to the top of the vegetative canopy and from molecular to global scales, through an iterative cycle of model-driven experimentation and observation dubbed MODEX. Considerable progress has been made toward achieving this overarching goal, but widely recognized is the fragmentation across projects and disciplines of the relevant modeling and simulation capabilities, observational and experimental data, analysis algorithms, and workflow tools. This fragmentation creates significant challenges for studying impacts and feedbacks in these complex multiscale systems. These challenges are further exacerbated by ongoing disruptive changes in high-performance computational architectures and the exponential growth in the types and volume of data that render obsolete the conventional approaches to software development and data management. Overcoming these challenges will require the development of a BER Climate and Environmental Sciences Division (CESD)–wide enabling cyberinfrastructure to support data management, cross-domain modeling, data analysis, and collaborative research.

To explore the potential for working groups to initiate and guide a more integrated and community-based cyberinfrastructure, BER held the ESS Workshop on Model-Data Integration: Modeling Frameworks, Data Management, and Scientific Workflows on April 30–May 1, 2015, following the ESS Principal Investigator Meeting in Potomac, Maryland. Participants included model developers, software engineers, and data management specialists from eight national laboratories, which represented a wide range of projects and programs from CESD as well as cross-cutting projects from DOE's Office of Advanced Scientific Computing Research (ASCR; both BER and ASCR are operated from within DOE's Office of Science). A series of plenary talks provided background information and clarified the workshop's three objectives: (1) develop requirements for this community-based cyberinfrastructure to ensure enhanced scientific productivity of the community as a whole; (2) identify challenges associated with developing this new cyberinfrastructure using a phased approach guided by project-driven use cases; and, given these requirements and challenges, (3) chart a path forward for ESS working groups to lead the phased development of the new cyberinfrastructure.

To address these objectives, three breakout sessions were organized and intermixed with lightning talks that provided additional information. The first breakout session discussed requirements and challenges for near-term development of the community-based cyberinfrastructure (Phase 1: 0 to 2 years), and was tasked with identifying initial capabilities that could be developed under ongoing BER-funded projects. Capabilities and use cases were identified in areas of data management, model interoperability and coupling, complex model and data workflow, and provenance capture.

The second breakout session focused on requirements and challenges for longer-term development (Phase 2: 2 to 5 years; Phase 3: 5 to 10 years) and split discussions into two subtopics. Subtopic one targeted multiphysics-multiscale process coupling. A key finding was the potential for community-based, flexible multiphysics and multiscale frameworks to enable sharing of capabilities across projects and scales to significantly enhance predictive understanding. Subtopic two examined model-data integration workflows and touched on issues surrounding the collection of model input data, model parameterization, initialization, and uncertainty quantification. This discussion identified several high-priority capabilities that would naturally be supported in a community-based cyberinfrastructure and significantly enhance scientific productivity, including metadata archiving, code sharing for parameterization, and modular parameter estimation and uncertainty quantification.

The third breakout session centered on operational issues associated with ESS working groups meeting the requirements and addressing the challenges identified in the first two breakout sessions. Three topics were identified for discussion: (1) working group governance and management, (2) setting of working group priorities and their relationships to existing ESS projects, and (3) licensing and intellectual property issues associated with the code and tools comprising the proposed community-based cyberinfrastructure. The first discussion identified governance as a critical factor in ensuring that working groups support the overall community and add value to existing ESS projects. The second discussion concluded that working group priorities should be set through input from the broader ESS community. Finally, the licensing and intellectual property discussion raised several important points relating to the established funding model, in which modeling capabilities are often considered a critical part of a team's competitive advantage. Specifically, shifting to a community-based cyberinfrastructure that enhances sharing of capabilities and accelerates the development of predictive understanding requires a research business model that acknowledges and even rewards these contributions. A complete solution to this complex and critical piece of community-based cyberinfrastructure does not yet exist, but the incredible growth of open-source scientific software provides a solid foundation upon which to build. In addition, other DOE offices (e.g., ASCR) and federal agencies (e.g., National Science Foundation) are facing these same challenges, creating collaboration and leveraging opportunities through the adoption of common policies and building of consensus across the broader scientific community on the approach and implementation.

Based on these workshop discussions, this report proposes a two-level structure for the CESD-ESS cyberinfrastructure working groups: an overarching executive committee and dynamically formed working groups to address specific topics and scope. Key topics identified for the initial set of working groups include data management, model-data integration, software engineering and interoperability, and community governance. The associated activities are well aligned with existing ESS projects and will benefit the community as a whole. This report recommends the launch of the ESS executive committee and formation of these four working groups.

1. Motivation and Vision for a Community Cyberinfrastructure

The Environmental System Science (ESS) activity of the Office of Biological and Environmental Research (BER) within the Department of Energy's (DOE) Office of Science seeks to advance a robust, predictive understanding of terrestrial environments, extending from bedrock to the top of the vegetative canopy and from molecular to global scales in support of DOE's energy and environment missions. Using an iterative approach of model-driven experimentation and observation (MODEX), interdisciplinary teams of scientists work to unravel the coupled physical, chemical, and biological processes that control the structure and functioning of terrestrial environments across vast spatial and temporal scales. State-of-science understanding is captured in conceptual theories and models and translated into a hierarchy of computational components used to predict the system's response to perturbations caused, for example, by changes in climate, land use or cover, or contaminant loading. Basic understanding of the system's structure and function is advanced through this iterative cycle of experimentation and observation by targeting key system components and processes suspected to most limit the predictive skill of the models.

The efficiency of this iterative cycle is critical to advancing the predictive understanding sought by ESS, but the current cyberinfrastructure used by ESS-supported scientists within BER's Climate and Environmental Sciences Division (CESD) is fragmented among various disciplines, organizations, and physical and temporal scales. This fragmentation leads to significant challenges, both logistical and fundamental. These challenges are exacerbated by ongoing disruptive changes in high-performance computational architectures and exponential growth in the types and volume of data that render obsolete the conventional approaches to software development and data management (U.S. DOE 2013). Widely recognized is that a more seamless and robust cyberinfrastructure that fully enables data management, cross-domain modeling, data analysis, and collaborative research is required to accelerate this iterative approach to scientific discovery (BERAC 2013).

Thus, there is an emerging consensus that developing a CESD-wide, open-source cyberinfrastructure (one that would contain data management, model development, data assimilation, and software engineering components) is valuable and now feasible. In addition, using interdisciplinary and interproject working groups to initiate and coordinate the cyberinfrastructure's development would foster integration of data management and model development across CESD programs (e.g., Climate and Earth System Modeling program) and provide an accessible focal point for collaboration and integration with other projects (U.S. DOE 2014). Within the United States, these projects could include the multiagency Earth System Grid Federation (ESGF), as well as the EarthCube (National Science Foundation) and the Community Earth System Model and Data Assimilation Research Testbed [CESM and DART, National Center for Atmospheric Research (NCAR)]. International projects could include the Data Integration and Analysis System (Japan).

The fragmented nature of the cyberinfrastructure challenge, along with traditional funding models, is driving the need for new research approaches that facilitate flexible integration of teams and capabilities across existing projects and programs. An emerging strategy is an agile and phased approach based on science-driven use cases (U.S. DOE 2015). In this setting, scientific questions are the key drivers and offer a dual view of this challenge. First, high-level science drivers provide a holistic view that can help prioritize and direct activities to ensure multiple projects and programs can share and benefit from targeted capability development. In traditional terms, this aspect of the strategy is *top down*. Second, scientific questions within a project or program can be used to define specific use cases, where development of new capabilities, or interoperability between existing capabilities, is critically needed. Because

these scientific questions are often complex and multifaceted, they naturally lead to a series of intermediate targets that can be addressed over time using a phased approach. A critical aspect of this phased approach is that at each step the capability or interoperability mechanism being developed will be useful to the project and the entire community. With scientists driving use cases in specific projects, and with phases spanning near- to long-term time frames, this aspect of the strategy is considered *bottom up*. Finally, science is naturally dynamic and iterative—as new understanding is gained through research, new questions and challenges emerge. For cyberinfrastructure, this progression brings (often rapidly) evolving requirements, which create significant challenges for managing software development and productivity. The software development methodologies that have emerged to embrace this dynamic environment are referred to as agile methods and are well suited to this application.

Together, this challenge and general approach motivated the recent ESS Workshop on Model-Data Integration: Modeling Frameworks, Data Management, and Scientific Workflows. This workshop had three primary objectives. (1) Develop requirements for a community-based cyberinfrastructure to ensure enhanced scientific productivity of the community as a whole. For example, participants discussed and characterized interoperability requirements for capabilities in data management, model development, and MODEX-driven integration. (2) Identify the core issues associated with developing this community-based cyberinfrastructure using a phased approach over the next decade. For this objective, participants expressed the key technical challenges as project-driven use cases and discussed and prioritized operational challenges, such as software licensing standards and funding models. (3) Given these requirements, operational challenges, and use cases, chart a path forward for the phased development of the community-based cyberinfrastructure to support ESS goals. This path would include both an operational structure and governance for interdisciplinary and interproject working groups, as well as several short-term goals aligned with current ESS projects.

2. Workshop Overview

The ESS Workshop on Model-Data Integration: Modeling Frameworks, Data Management, and Scientific Workflows was held at the Bolger Center in Potomac, Maryland, on April 30–May 1, 2015, following the ESS Principal Investigator Meeting (see Appendix A. Workshop Agenda, p. 10). Participants included nearly 40 model developers, software engineers, and data management professionals from eight national laboratories (see Appendix B. Workshop Organizers and Participants, p. 12), representing the ESS Scientific Focus Areas (SFAs), Next-Generation Ecosystem Experiments projects (NGEE–Arctic and NGEE–Tropics), and Interoperable Design of Extreme-scale Application Software (IDEAS) project. Multiple program managers from BER and DOE's Office of Advanced Scientific Computing Research (ASCR) also attended the workshop.

A series of plenary talks kicked off the workshop, with the first two framing the model-data integration challenges from the perspective of ESS and CESD programs. In particular, these talks highlighted the fragmented state of current capabilities and tools critical to model-data integration and raised the potential for interdisciplinary and interproject working groups to address this challenge. The second two talks provided examples of community-based tools and approaches from the IDEAS and ESGF projects that the working groups could implement.

The plenary session concluded with three talks connecting the science drivers in specific projects, first to design requirements for an integrated, community-based cyberinfrastructure and then to use cases that would support a phased approach to this development. The first of these talks focused on Terrestrial Ecosystem Science (TES) projects (NGEE–Arctic and NGEE–Tropics), the second on Subsurface Biogeochemical Research (SBR) projects [Lawrence Berkeley National Laboratory (LBNL) and Pacific Northwest National Laboratory (PNNL) SFAs], and the third on data management across both TES and SBR, which make up the ESS activity. Although each of these projects has distinct scientific goals, common themes and needs were identified and common design requirements emerged, including

- Coupling multiscale (temporal and spatial) observations and multiscale models.
- Formally coupling models that exist in different domains (e.g., plants and soil, soil and atmosphere, and biogeochemistry and hydrology).
- Performing quantitative and formalized uncertainty quantifications.
- Leveraging (as much as is possible) existing code capabilities.
- Rigorous, but rapid testing and validation of model-data integration capabilities.
- Increasing scientific productivity (compared to current approaches) through open and interoperable capabilities provided by a community-based cyberinfrastructure.

Use cases that would drive the development of capabilities to meet these design requirements include

- Quantitative modeling of hot spots and hot moments (IDEAS; LBNL and PNNL SFAs).
- Tight coupling of plant hydraulics from water table to canopy (NGEE–Tropics).
- Componentization of existing modeling capabilities through a well-defined interface to enable sharing of capabilities between projects and codes.
- Development of a community-accepted modular modeling approach with flexible data abstractions and well-defined interfaces, along with demonstrated feasibility for use with existing models.
- Development of a common data management framework and associated modular tools across SBR. Such a data management framework would borrow from existing efforts.

- A well-architected data-model "linkage" to obtain data from distributed data stores (by well-defined interfaces and abstracted queries) and to return (as one of the outputs) information on data needs required for model enhancement.

Setting the stage for breakout sessions that addressed near- and longer-term goals for the community-based CESD-ESS cyberinfrastructure (see Appendix C. Breakout Session Summaries, p. 14) were the above-mentioned requirements and use cases, in concert with lightning talks on computational plans and resources at the individual national laboratories (see Appendices D and E, p. 19 and p. 31, respectively, for descriptions of national laboratory computational resources and plans, as well as computational trends). The first breakout session focused on the near term (Phase 1: 0 to 2 years), while the second breakout session focused on the mid to long term (Phase 2: 3 to 5 years; Phase 3: 5 to 10 years). Both of these breakouts identified requirements and challenges for developing the community-based cyberinfrastructure within their respective time frames (phases), and the results of these discussions are summarized in Appendices C.1 and C.2, p. 14 and p. 15, respectively. The third breakout session supported the workshop's third objective and focused on operational issues associated with ESS working groups meeting these requirements and challenges to develop and support the envisioned community-based cyberinfrastructure. These discussion highlights are captured in Appendix C.3, p. 17. Critical session findings include: (1) **governance is a critical factor in ensuring that the working groups support the overall community and add value to existing ESS projects**, (2) working group priorities should be set with input from the broader ESS community, and (3) shifting to a community-based cyberinfrastructure that enhances sharing of capabilities and accelerates the development of predictive understanding **requires a research business model that acknowledges and even rewards these contributions**.

Based on feedback from these three breakout sessions, a working group organizational structure is presented in the next section (see Section 3.1, p. 5), followed by a series of next steps for four key working group topic areas (see Section 3.2, p. 6).

3. Workshop Outcomes

Collectively, the three breakout sessions addressed the workshop's objectives and have led to the following proposed working group structure and governance and a set of critical topics and next steps.

3.1 Proposed Working Group Structure and Governance

Building on the workshop discussions summarized in Appendix C.3, p. 17, an initial concept for an agile, two-level structure is proposed for the CESD-ESS cyberinfrastructure working groups. This structure comprises an overarching executive committee and dynamically formed working groups, which will develop the cyberinfrastructure capabilities through a phased, use case–driven approach. Specifically, the envisioned executive committee structure borrows from the ESGF governance strucure. The executive committee is a flat organization whose members include principal investigators (PIs) from the major ESS projects and a smaller number of "passionate individuals" chosen to broadly represent data, observations and experimentation, model development, and software engineering. Additionally, a steering committee consisting of program managers from DOE and other interested agencies will work with the executive committee to help identify and prioritize the topics that the working groups address. Executive committee membership will approach nine to 12 people (including six to 10 PIs and three to five additional members), with non-PI members rotating on 3 year terms. Responsible for maintaining representation across the range of model-data integration expertise, the executive committee will identify and invite new members to replace members rotating off. A chair and co-chair will be identified from within the executive committee membership and be responsible for organizing twice-annual meetings. The chair and co-chair will develop meeting agendas with input from the entire executive committee and extend invitations to guest speakers. DOE program management will provide meeting logistical support. Meetings will be conducted with DOE and other agency program management in attendance, providing an opportunity for program managers to describe current program direction and to ask questions of and field questions from executive committee members.

Topics within the executive committee's charge will be explored in detail by the working groups. The executive committee can vote to initiate the formation of working groups, and working groups will be free to seek input as needed from outside the executive committee. Each working group will have a well-defined charge, set its own meeting schedule, define a clear set of deliverables, and dissolve as its work is completed. Multiple working groups can be run in parallel, with their efforts reported to the entire executive committee at regular meetings. Support for needed working group travel and workshops outside the groups' regular meeting schedule will be provided after approval by project or program leadership.

Initial population of the executive committee will be accomplished by DOE program management from the list of ESS project PIs. Those PIs then will designate the remaining membership. The chairs will be selected through a nomination and voting procedure. Project PIs will be allowed to name a delegate to replace them on the executive committee, either for particular meetings or as long-term delegates. Non-PI members of the initial executive committee will be assigned terms of 2, 3, or 4 years to ensure a balance of fresh perspectives with historical knowledge and experience.

Once the initial executive committee is formed, the set of tasks described in the next section will be prioritized and grouped under topics. At this stage, higher priority will be given to tasks with greater payoff potential in shorter time frames, and largely within the scope of existing projects. Subsequently, a small number of subgroups (two to four) will be formed to address the highest priority topics.

3.2 Next Steps

This report strives to capture the essence of the insightful workshop discussions that took place, as well as the key challenges in model-data integration, where interproject working groups can significantly enhance the scientific productivity of the community as a whole. Focusing on this objective, workshop participants considered a phased approach to identify near-term activities aligned with current projects that, therefore, can be pursued within existing project funding. Although these activities may not have been called out explicitly in the past, their pursuit can leverage multiple projects because their completion will benefit multiple projects.

Workshop participants identified the following four potential working group topics, along with their associated near-term activities:

Data Management

- Define a common standard for using digital object identifiers (DOIs) or similar digital identifiers to make data publishable and citable.
- Develop best practices and use case–driven templates for data archiving, beginning with observational data and extending to all supporting data and workflow information that supports validation as defined in the data management plans.

Model-Data Integration

- Survey the ESS community to evaluate model-data integration workflows and documentation procedures and develop best practices guidance.
- Identify existing resources or tools that could help standardize workflows and workflow documentation.

Software Engineering and Interoperability

- Assess data formats used in existing codes, develop a taxonomy or classification scheme, and identify data specification needs.
- Create community data specifications for selected process models.

Community Governance

- Determine areas where ESS can provide critical guidance and community support (e.g., clarification of code sharing and open-source licensing requirements and formal recognition of software development expertise and productivity).
- Develop guidelines for community-oriented services or standards and the potential need for ESS support of these services.
- Identify working group members to serve as champions who will advocate for the adoption of ESS tools, processes, and workflows by the greater community.

This initial set of activities targets a 6-month trial phase for each working group. If a working group is productive and members feel it should continue to tackle new scopes, then additional activities can be added. In contrast, if a working group has completed its tasks or is not being productive, it can be dissolved.

To move forward with the new CESD-ESS cyberinfrastructure, an executive committee should be established first, as outlined previously in Section 3.1, which then would form an initial set of working groups. Next, each working group, in coordination with program PIs, can formalize the scope and deliverables for the group's activities by taking a use case–driven approach. Based on this strategy, the following timeline for these next steps is proposed:

July 2015

- Submit a draft workshop report to DOE program managers for initial feedback and finalize the draft to share with workshop participants.

August 2015

- Distribute a draft workshop report to the participants and DOE program managers for final review and comment.
- Discussion between program PIs and DOE program managers to obtain commitment for participation in working group activities.

September 2015

- Complete updates to the report based on feedback and finalize for publication.
- Establish the PI-based part of the executive committee.
- Select one or two working group topics and outline potential working group membership.

November 2015

- Complete creation of the executive committee.
- Launch at least two working groups.

Quarterly

- Report to the executive committee on working group task progress.

2016 ESS PI Meeting

- Report on working group task status**.**

References

BERAC. 2013. *BER Virtual Laboratory: Innovative Framework for Biological and Environmental Grand Challenges; A Report from the Biological and Environmental Research Advisory Committee,* DOE/SC-0156. Biological and Environmental Research Advisory Committee, U.S. Department of Energy Office of Science, Washington, D.C. (http://science.energy.gov/ber/berac/reports/).

U.S. DOE. 2015. *Building Virtual Ecosystems: Computational Challenges for Mechanistic Modeling of Terrestrial Environments; Workshop Report*, DOE/SC-0171. U.S. Department of Energy Office of Science, Washington, D.C. (http://doesbr.org/BuildingVirtualEcosystems/).

U.S. DOE. 2014. *Data-Model Needs for Belowground Ecology: A Summary Report from the Terrestrial Ecosystem Science (TES) Mini-Workshop*. U.S. Department of Energy Office of Science (http://tes.science.energy.gov/workshops/).

U.S. DOE. 2013. *Extreme-Scale Scientific Application Software Productivity: Harnessing the Full Capability of Extreme-Scale Computing.* White paper prepared for the U.S. Department of Energy Office of Advanced Scientific Computing Research. (www.orau.gov/swproductivity2014/ExtremeScaleScientificApplicationSoftwareProductivity2013.pdf)

U.S. DOE. 2002. "Policy Guidance – OSS License Release of Software Developed with ASC and OASCR Funding," Memo from the U.S. Department of Energy Offices of Advanced Simulation and Computing and Advanced Scientific Computing Research (http://science.energy.gov/~/media/ascr/pdf/research/docs/Doe_lab_developed_software_policy.pdf).

Appendices

Appendix A. Workshop Agenda

Environmental System Science (ESS) Workshop on Model-Data Integration: Modeling Frameworks, Data Management, and Scientific Workflows

April 30 – May 1, 2015

Thursday, April 30, 2015

8:30 a.m. – 8:45 a.m. **Welcome: Goals for ESS Working Group and Workshop**
David Lesmes, Department of Energy Office of Biological and Environmental Research (BER) Climate and Environmental Sciences Division (CESD)

8:45 a.m. – 9:00 a.m. **CESD Integrated Data System and Workflow**
Jay Hnilo, BER CESD

9:00 a.m. – 9:15 a.m. **Scientific Software Engineering and Productivity: IDEAS Project Overview and Use Cases**
David Moulton, Los Alamos National Laboratory (LANL)

9:15 a.m. – 9:30 a.m. **Community Tools to Facilitate Model-Data Integration Across Scales**
Dean Williams, Lawrence Livermore National Laboratory (LLNL)

9:30 a.m. – 9:45 a.m. **Open Discussion: Goals, Design Requirements, and Implementation**

9:45 a.m. – 10:00 a.m. **BREAK**

ESS Science Drivers → Design Requirements → Use Cases

10:00 a.m. – 10:15 a.m. **Model-Data Integration Challenges and Opportunities: Terrestrial Ecosystem Science (TES)**
Peter Thornton, Oak Ridge National Laboratory (ORNL); Charlie Koven, Lawrence Berkeley National Laboratory (LBNL)

10:15 a.m. – 10:30 a.m. **Model-Data Integration Challenges and Opportunities: Subsurface Biogeochemical Research (SBR)**
Carl Steefel, LBNL; Tim Scheibe, Pacific Northwest National Laboratory (PNNL)

10:30 a.m. – 10:45 a.m. **Data Management Challenges and Opportunities: TES and SBR**
Deb Agarwal, LBNL; Tom Boden, ORNL; Roelof Versteeg, Subsurface Insights, Inc.

10:45 a.m. – 11:00 a.m. **Open Discussion**

11:00 a.m. - 11:15 a.m. **Introduction to Breakout 1 (Working Lunch)**

11:15 a.m. – 1:15 p.m. **Breakout 1: Science Drivers → Design Requirements → Use Cases**
Discuss and refine overarching goals and principles for implementation and continue gathering and refining design requirements and use cases that can be implemented in ***Phase 1 (0 to 2 years)***

Session 1A: Modelers, Data Management
Facilitator: Deb Agarwal, Scribe: Tom Boden, Presenter: Lara Kueppers

Session 1B: Model Interoperability and Coupling Interface Definitions
Facilitator: Tim Scheibe, Scribe: Ethan Coon, Presenter: David Moulton

Session 1C: Complex Model-Data Workflow and Provenance Capture
Facilitator: Peter Thornton, Scribe: Roelof Versteeg, Presenter: Shawn Serbin

1:15 p.m. – 1:45 p.m. **Reports from Breakout 1: Sessions 1A, 1B, and 1C**

1:45 p.m. – 2:00 p.m. **DISCUSSION**

2:00 p.m. – 2:30 p.m. **Lightning Talks About Computational Plans and Resources that Support and Inform ESS Projects and Programs Within CESD**
(Note: Assignment is to develop 2-page lab summaries)

- *Peter Thornton, ORNL*
- *Carl Steefel, LBNL*
- *Kerstin Kleese van Dam, PNNL*
- *Ethan Coon, LANL*
- *Umakant Mishra, Argonne National Laboratory (ANL)*

2:30 p.m. – 2:45 p.m. **Software Engineering Tools and Methodologies for Community Code Development: What Support Can the IDEAS Project Provide?**
David Bernholdt, ORNL; Lois Curfman McInnes, ANL; Hans Johansen, LBNL

2:45 p.m. – 3:00 p.m. **DISCUSSION**

3:00 p.m. – 3:15 p.m. **Introduction to Breakout 2**

3:15 p.m. – 3:30 p.m. **BREAK**

3:30 p.m. – 5:30 p.m. **Breakout 2: Prioritization of ESS Software-Infrastructure Needs: Developing an Integrated Software Ecosystem to Facilitate and Accelerate Model-Data Integration and Scientific Productivity**
(Develop a 10 year plan, based on a three-phased approach—0 to 2 years, 2 to 5 years, 5 to 10 years—and general use cases)

Session 2A: Multiphysics, Multiscale Process Coupling
Facilitator: Carl Steefel, Scribe: Jeff Johnson, Presenter: Charlie Koven

Session 2B: Model-Data Integration Workflow
Facilitator: Haruko Wainwright, Scribe: Tom Boden, Presenter: Xingyuan Chen

5:30 p.m. – 6:00 p.m. **Status Report from Breakout 2: Sessions 2A and 2B**
(Breakout discussions could continue over dinner)

Friday, May 1, 2015

8:30 a.m. – 8:45 a.m. **Lightning Talks About Computational Trends, Developments, Challenges and Opportunities: Informing ESS Projects and Programs Within CESD**
(Note: Assignment is to develop a 1- to 2-page white paper on topic)

- *Lois Curfman McInnes, ANL*
- *David Bernholdt, ORNL*
- *Deb Agarwal, LBNL*
- *David Moulton (for Pat McCormick), LANL*
- *Kerstin Kleese van Dam, PNNL*

8:45 a.m. – 9:00 a.m. **DISCUSSION About Lightning Talks**

9:00 a.m. – 9:45 a.m. **Review of Day 1 and Open Discussion About Plans for Developing the ESS Integrated Software Ecosystem Using a Phased Approach**

9:45 a.m. – 10:00 a.m. **Introduction to Breakout 3**

10:00 a.m. – 10:15 a.m. **Management and Governance of Community Modeling Frameworks**
Metrics of Success for Community Modeling Frameworks
Glenn Hammond, Sandia National Laboratories

Licensing Options and Considerations for Open-Source Community Codes
David Moulton, LANL; Tim Johnson, PNNL

Management and Governance of the Earth System Grid Federation (ESGF): An Illustrative Example
Dean Williams, LLNL; Jay Hnilo, BER CESD

10:15 a.m. – 10:30 a.m. **BREAK**

10:30 a.m. – 11:30 a.m. **Breakout 3: ESS Working Group – Management, Licensing, and Governance**
(Organization of the working group – how will it work?)

Session 3A: Software Engineering Team
Facilitator: David Moulton, Scribe: Tim Scheibe, Presenter: Tim Johnson

Appendix B. Workshop Organizers and Participants

ORGANIZING COMMITTEE

Workshop Chair: David Moulton
Los Alamos National Laboratory

Software Engineering

David Moulton
Los Alamos National Laboratory

Dean Williams
Lawrence Livermore National Laboratory

Data Management

Deb Agarwal
Lawrence Berkeley National Laboratory

Tom Boden
Oak Ridge National Laboratory

Roelof Versteeg
Subsurface Insights, Inc.

Model Development

Charlie Koven
Lawrence Berkeley National Laboratory

Tim Scheibe
Pacific Northwest National Laboratory

Carl Steefel
Lawrence Berkeley National Laboratory

Peter Thornton
Oak Ridge National Laboratory

DOE BER ORGANIZERS

Climate and Environmental Sciences Division

David Lesmes
David.Lesmes@science.doe.gov

Justin Hnilo
Justin.Hnilo@science.doe.gov

DOE PROGRAM MANAGER ATTENDANCE

Office of Biological and Environmental Research

Paul Bayer
Subsurface Biogeochemical Research

Gary Geernaert
Climate and Environmental Sciences Division Director

Renu Joseph
Regional and Global Climate Modeling

Mike Kuperberg
Terrestrial Ecosystem Science

Sally McFarlane
ARM Climate Research Facility

Shaima Nasiri
Atmospheric System Research

Rick Petty
ARM Climate Research Facility

Dan Stover
Terrestrial Ecosystem Science

Sharlene Weatherwax
BER Associate Director

Office of Advanced Scientific Computing Research

Richard Carlson
Collaboratories/Middleware

Thomas Ndousse-Fetter
Network Research

PARTICIPANTS

Software Development

David Bernholdt
Oak Ridge National Laboratory

Ethan Coon
Los Alamos National Laboratory

Lois Curfman McInnes
Argonne National Laboratory

Glenn Hammond
Sandia National Laboratories

Jeff Johnson
Lawrence Berkeley National Laboratory

Kerstin Kleese van Dam
Pacific Northwest National Laboratory

Sergi Molins
Lawrence Berkeley National Laboratory

Dali Wang
Oak Ridge National Laboratory

Data Management

Xingyuan Chen
Pacific Northwest National Laboratory

Les Hook
Oak Ridge National Laboratory

Lara Kueppers
Lawrence Berkeley National Laboratory

Umakant Mishra
Argonne National Laboratory

Shawn Serbin
Brookhaven National Laboratory

Haruko Wainwright
Lawrence Berkeley National Laboratory

Model Development

Gautam Bisht
Lawrence Berkeley National Laboratory

Eoin Brodie
Lawrence Berkeley National Laboratory

Forrest Hoffman
Oak Ridge National Laboratory

Maoyi Huang
Pacific Northwest National Laboratory

Jitendra Kumar
Oak Ridge National Laboratory

Melanie Mayes
Oak Ridge National Laboratory

Scott Painter
Oak Ridge National Laboratory

Daniel Ricciuto
Oak Ridge National Laboratory

Bill Riley
Lawrence Berkeley National Laboratory

Joel Rowland
Los Alamos National Laboratory

Hyun-Seob Song
Pacific Northwest National Laboratory

Stan Wullschleger
Oak Ridge National Laboratory

Chonggang Xu
Los Alamos National Laboratory

Appendix C. Breakout Session Summaries

C.1 Breakout Session 1: (Short Term) Science Drivers → Design Requirements → Use Cases

The objective of this breakout session was to identify a small number of Phase 1 (immediate, 0 to 2 years) efforts (based on science drivers) that would result in initial capabilities in a shared (community-based) cyberinfrastructure. The general concept, as explained by Department of Energy (DOE) program managers, is that in Phase 1, working group–related efforts would be undertaken by the Environmental System Science (ESS) community as part of ongoing major Office of Biological and Environmental Research (BER)–funded projects, whereas in subsequent phases (Phase 2: mid term, 3 to 5 years; and Phase 3: long term, 6 to 10 years), dedicated cyberinfrastructure funding would most likely be available.

The criteria for Phase 1 efforts are that they (1) should benefit multiple projects, (2) be synergistic with ongoing cyberinfrastructure efforts (e.g., model-data integration and data management), and (3) seem achievable within funding and operational constraints.

Breakout Session 1 was organized into three groups:

- 1A: Data Management
- 1B: Model Interoperability and Coupling Interface Definitions
- 1C: Complex Model-Data Workflow and Provenance Capture

Group 1A: Data Management

This group discussed a wide range of data management activities that are needed by most ESS projects and that offer opportunities to accomplish collaborative capabilities in the 0 to 2 year time frame. The first use case discussed was a data archiving/publishing and citation capability. Data publication at the time of paper publication is a DOE requirement that all ESS scientists must adhere to (as of October 1, 2014), but it is recognized that there are substantial challenges in meeting this requirement. Coordinated efforts provide an opportunity to define a common standard for using digital object identifiers (DOIs) or similar digital identifiers to make data publishable and citable. The idea is to define a standard for developing data packages to archive data in logical groups, including defining a common format and metadata standard for archiving the data associated with a publication. Agreement on a common template for metadata information associated with a data package would be valuable, because it would enable development of data access portals that provide access to data across multiple projects. Also discussed was the potential for collaboration in defining common data collection templates.

The group agreed that archiving is exceedingly important in all projects, but the range of data that needs to be archived is still somewhat open for discussion. The maximum range includes all supporting data, software, models, model outputs, scripts, and workflows. One idea discussed was to consider defining a minimum requirement, which individual projects could choose to exceed based on their needs. Another idea was the development of data collection template generators that would provide standard metadata fields and data reporting fields to simplify data archiving and use.

Group 1B: Model Interoperability and Coupling Interface Definitions

This group focused on definitions of model interfaces for model coupling and data standards for model input and output to provide the initial foundation for the long-term goal of model operability across the ESS community. Discussion centered on a number of issues related to model coupling—in particular, tight versus loose coupling, data transfer through files or in memory, and data specifications potentially leading to an application programming interface (API). A number of existing tools and capabilities were mentioned and briefly discussed, including the process kernels and multiprocess coupler in the Alquimia data mediator and interface library for linking to geochemistry engines (Amanzi/Arcos) and the process couplers in the suites of codes for the Accelerated Climate Modeling for Energy (ACME) model coupling toolkit (MCT) and the massively parallel reactive flow and transport model for describing surface and subsurface processes (PFLOTRAN) physics model coupler (PMC). Also noted was a number of emerging tools for multiscale coupling, such as the MML (Multiscale Modeling Language) and MAPPER (Multiscale Applications on European e-Infrastructures) frameworks, capabilities of which the group was largely unaware.

Based on the general discussion, a sequence of five potential short- to medium-term activities was identified. The proposed community resources and associated perceived potential impact are (in temporal order):

- Assess data formats used in existing codes and develop a taxonomy or classification scheme [e.g., partial differential equations (PDEs), ordinary differential equations (ODEs), differential algebraic equations (DAEs), stochastic differential equations (SDEs), and reduced-order models (ROMs)]. *Impact*: Provides a foundation for data specification, a necessary step for community adoption.
- Identify needs for data specification(s) and work through use cases (e.g., flow and transport, rhizosphere biogeochemical dynamics, plant hydraulics, and metabolic modeling). *Impact*: Engages the community, developing acceptance.
- Create community data specification(s) for selected process models. *Impact*: Facilitates data exchange and sharing of capabilities, empowering people to contribute to the software ecosystem.
- Develop APIs for data mediators (interoperable, extensible, and agnostic). *Impact*: Eases adoption by developing familiarity with software engineering practices and encouraging best practices.
- Develop APIs for model components and couplers. *Impact*: Enables flexible use and coupling of components as well as interoperability and sharing across projects.

The first three activities were identified as being feasible in the short term (0 to 2 years) with limited resource investment. The latter two are longer-term goals that would be enabled by the initial three steps.

Group 1C: Complex Model-Data Workflow and Provenance Capture

This group was charged with identifying challenges in the area of model-data workflow and provenance capture that could be addressed in the 0 to 2 year time frame by an ESS working group. The intent was to identify initial steps toward a long-term, broadly defined goal of better tools and practices for model-data integration.

The group self-identified as six modelers, three software engineers, six observationalists and experimentalists or data management professionals, and three synthesis specialists, but the categories are not mutually exclusive; that is, many identified in more than one category.

Given the short time frame of interest (0 to 2 years), the group focused on issues of transparency and traceability in model results including model-data integration activities. The group recognized that traceability is an initial step in the direction of the more stringent requirement of reproducibility. Although a formal definition for transparency or traceability was not discussed or suggested, transparency and traceability were used interchangeably to refer to a process that enables someone with a relevant technical background, but who was not involved with the work, to assess the authenticity and scientific validity of the results. A great deal of variability seems to exist among the various Scientific Focus Areas (SFAs) and multilaboratory projects regarding how key information is recorded such as software versions, computational platforms, model parameter values, and data sources.

The following near-term actions were identified:

- Survey the ESS community to determine which model-data workflows are used and how the steps and intermediate results are documented.
- Summarize current practices in documenting workflows, intermediate results, and final results.
- Identify existing resources that could help standardize workflows and workflow documentation (e.g., assignment of DOIs to workflows).
- Make recommendations on best practices.
- Present the results of these activities at the next ESS principal investigator meeting or in a white paper.

C.2 Breakout Session 2: (Longer Term) Prioritization of ESS Software-Infrastructure Needs

Group 2A: Multiphysics, Multiscale Process Coupling

This group focused on multiscale systems with potentially differing physics (or physics representations) as well as data types. Changing scales often involves a shifting importance for various processes as well as the extent to which they are loosely or tightly coupled to other representations (e.g., vegetation responding to local hydrologic and biogeochemical hotspots that behave differently from other often much larger portions of the domain).

One example briefly explored was a multiscale system in which biogeochemical cycling was strongly affected by local (within 10 to 30 meters) biogeochemical hotspots and by potentially short-lived hydrologic transients (e.g., rainfall events and snow melt). In the most challenging case, the biogeochemical hotspots are repeated down the length of the system, as in the meanders of

the lower East River watershed (Gunnison County, Colorado) that is a focus of Lawrence Berkeley National Laboratory's Genomes to Watershed SFA and Use Case 1 of the Interoperable Design of Extreme-scale Application Software (IDEAS) project. In this system, the boundary conditions for downstream hotspots are affected by upstream behavior, which makes all the boundary conditions for local hotspots time dependent. A second case involves the tight coupling between soil and plant hydraulics, which easily could be extended to nutrient uptake and biogeochemical cycling. The plant (leaf, stem, and root) and soil systems are coupled in a first-order fashion through their hydraulics, and a mathematical coupled system can be envisioned in which all process components are solved simultaneously. This implies a framework capable of handling coupling terms between individual processes, typically off-diagonal Jacobian blocks as in Fig. 2A-1. Process-Level Models for Subsystems, this page.

Group 2B: Model-Data Integration Workflow

This group focused on model-data integration workflows, which include the key steps of (1) extracting and processing necessary information from databases, (2) parameterizing and initializing models, (3) validating models, (4) performing uncertainty quantification (UQ) of model predictions, and (5) providing feedback from models to data acquisitions through experimental designs and data-worth analysis. A long-term (2 to 10 year) goal of this working group envisions the development of software or software ecosystems to facilitate and automate these key steps and connect them in a more seamless manner from databases to numerical simulators. Particularly recognized was the importance of being able to document and archive the workflow for traceability, transparency, and reproducibility. The following functionalities were identified as priority items:

- Metadata archiving.
- Data discovery capability to find available datasets in a prescribed model domain.
- Code sharing for data processing to parameterize models (e.g., interpolation).
- Workflow tracking for traceability, transparency, and reproducibility.
- Benchmarking and validation.
- Experimental design and data-worth analysis.

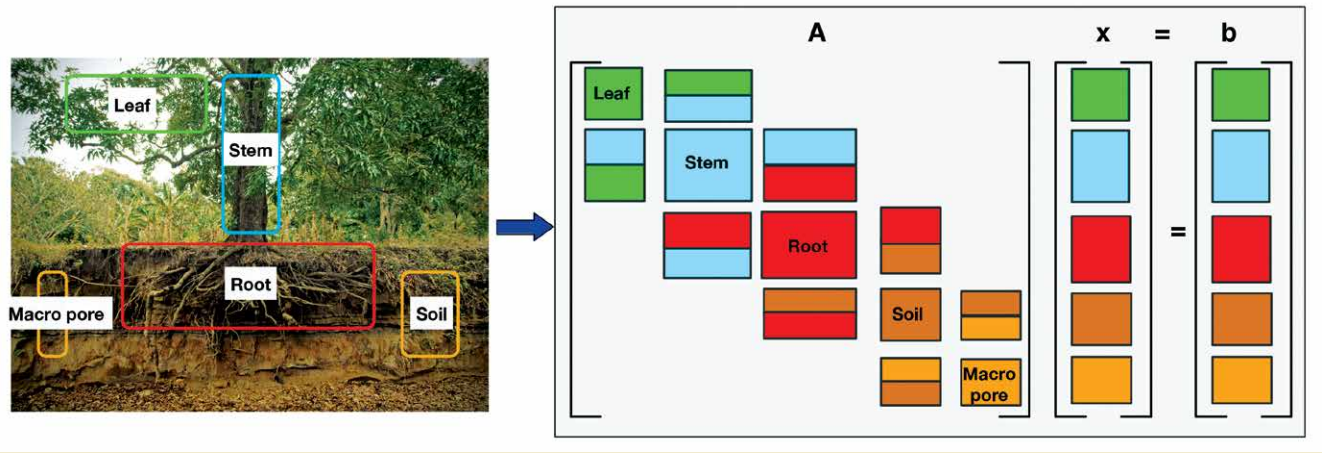


Fig. 2A-1. Process-Level Models for Subsystems. Process-level models are first developed for subsystems over limited spatial and temporal scales and then enhanced through model-data integration. Subsystem examples are shown in the left figure. These subsystems most often are studied independently on different projects through the development of unique modeling and simulation capabilities. However, exploring process coupling and large-scale system behavior cooperatively requires a flexible framework that can couple both model components (as suggested by the Jacobian matrix on the right). To support this flexibility, these components must present well-defined and open interfaces, adhere to data interoperability requirements, and be openly shared among projects. Using a phased approach, the proposed ESS working groups can address these requirements as they develop a community-based cyberinfrastructure that enhances these important collaborations and multiscale simulation and model-data integration capabilities. [Images courtesy Lawrence Berkeley National Laboratory]

- Modeling capabilities to inform the potential sources and impacts of data uncertainty.
- Downscaling and upscaling capabilities.
- Integrated and modular parameter estimation (PE) and UQ capabilities coupled with complex ecosystem and subsurface simulators. Modularity enables users to calibrate and test individual components [e.g., a vegetation component in the Community Land Model (CLM)].

These last two items are particularly important for accommodating next-generation multiscale, multiphysics simulators of terrestrial ecosystems. Also recognized is that more investment is needed, not only in data management and numerical simulators, but also in methodology and software development for model-data integration (including PE and UQ software).

C.3 Breakout Session 3: Governance and Management, Priority Setting, and Licensing and Intellectual Property

Given the near-term use cases and long-term vision established in the first two breakout sessions, this breakout was tasked with exploring operational issues associated with both the ESS programs and the proposed interdisciplinary working groups. For ESS programs, key questions revolve around the business model for scientific research using the model-driven experimentation and observation (MODEX) approach and the corresponding role and value of an integrated and shared cyberinfrastructure. In turn, these issues raise questions about how interdisciplinary working groups can be designed to foster collaboration, capability sharing, and community-based approaches that increase the community's overall scientific productivity.

To help focus the discussions on these operational issues and ensure that recommendations could be formulated on a path forward, three topics were identified:

- **Governance and Management:** How should the working groups be organized and managed?
- **Priority Setting:** Who decides what will be done within the working groups?
- **Licensing and Intellectual Property (IP):** How should ESS-related code be handled from a licensing and IP perspective?

Three talks were used to introduce these topics and provide real-world perspectives. In the first one, "Metrics of Success for Community Modeling Frameworks," Glenn Hammond presented his experience with PFLOTRAN on how open-source codes are developed and used and how valuable data about their use can be collected and analyzed. The second talk by David Moulton and Tim Johnson, "Licensing Options and Considerations for Open-Source Community Codes," addressed key issues facing developers and the support needed from funding sponsors. Finally, Dean Williams and Jay Hnilo discussed the governance of the Earth System Grid Federation, a key component in the community tools serving the climate science community.

Breakout Session 3 was then organized into the three topic areas. Each group primarily focused its discussion and report on its respective topic but was encouraged to discuss and comment on all three topics. In addition, participants were asked to fill out an online questionnaire related to this breakout during the workshop. The response rate was 60%.

Group 3A: Governance and Management

Governance was deemed a critical factor in ensuring that the working groups support the overall community and add value to existing ESS projects. Many participants felt this balance could be achieved with a distributed governance model, in which subgroups have significant autonomy, possibly feeding an overarching leadership council or group. There was consensus that the group of people involved in governance should be passionate about the community and include senior and junior scientists, large project leads, program managers, and big picture people. Guidelines on percent representation need to be developed, but no conclusions were drawn at the workshop. In addition, participants felt that stakeholders from the broader ESS community (not just developers) should be involved to ensure that their needs are captured and that they are aware of which capabilities are covered by each working group's activities. Similarly, participants acknowledged that interactions with other BER efforts (e.g., ACME) and across DOE's Office of Science (e.g., Scientific Discovery through Advanced Computing or SciDAC) are important.

There was general acceptance that no explicit funding is available for working group activities or community-oriented services at this time. However, as near-term deliverables are completed, participants noted the importance of **evolving a new ESS business model that acknowledges and rewards these contributions**. This new business model is necessary for BER to provide explicit funding for the cyberinfrastructure activities targeted by the working groups and ensure the

expected integration and synergy with existing programs. Such a business model also is a key factor in ensuring the health and productivity of the community as a whole as it establishes metrics beyond journal publications for scientists and program managers to gauge impact and value.

Group 3B: Priority Setting

Priorities and goals for the working groups need to balance narrower, short-term opportunities that are more closely tied to ongoing projects and work scope with the long-term vision of a community-based cyberinfrastructure. To strike this balance and best serve the community, most participants felt that **priorities for working group activities should be set by the broader ESS community** as opposed to ESS project leads or BER program managers. In addition, the consensus was that priorities should be science driven and supported by project-related use cases. This approach is preferable because it focuses activities on a series of impactful, incremental advances rather than large complex deliverables that cannot be easily managed or evaluated.

Group 3C: Licensing and IP

The general consensus was that ESS should provide open-source capabilities for the community to use, enhance, and customize. Because a number of technical issues exist relating to copyright, intellectual property, and compatibility of open-source licenses, the consensus was that these issues could be resolved most efficiently through direction from the program office—specifically, **policy- or memo-based guidance**, with reasonable freedom and flexibility for projects and including a clear statement of any exceptions as well as preferences and their benefits for the community [e.g., see the joint DOE Offices of Advanced Simulation and Computing and Advanced Scientific Computing Research memo (U.S. DOE 2002)]. Finally, with the growing number of sophisticated open-source software development and project management tools (e.g., github, gitlab, and bitbucket), the group recognized the need for best practices guidance and potential benefit of a community server to host multiple projects and provide uniformity and transparency.

Appendix D. Computational Resources and Plans in Support of Environmental System Science

Prior to the workshop, each of the national laboratories supported by the Department of Energy's (DOE) Office of Biological and Environmental Research (BER) was asked to provide a two-page description of capabilities and visions regarding the proposed cyberinfrastructure. Although the scope was left to the authors, the request suggested that the description should (1) highlight capabilities currently used within BER's Climate and Environmental Sciences Division (CESD) as well as those that may be used in the future, (2) note which of these capabilities currently are proprietary and which are open, and (3) comment on plans or developments that would make these capabilities more open to the community or part of a future software ecosystem.

Each of these descriptions is included in the following sections.

D.1 Argonne National Laboratory

Contributors: U. Mishra, B. A. Drewniak, Z. Fan, R. Jacob, J. D. Jastrow, K. M. Kemner, V. R. Kotamarthi, R. Matamala, L. C. McInnes, F. Meyer, and R. B. Ross

Argonne National Laboratory (ANL) through its various Earth System Science (ESS) and laboratory-supported activities has a number of projects contributing to the development of new techniques and computational tools that address specific needs of BER science. For example, the Argonne Leadership Computing Facility (ALCF) and the Mathematics and Computer Science (MCS) divisions have a large number of computer resources, software development research, and computing application capabilities (e.g., Mira, visualization clusters, data networking, and software). Furthermore, ANL conducts ESS-relevant research (bedrock to top of the boundary layer) through (1) Scientific Focus Areas [SFAs, including Accelerated Climate Modeling for Energy (ACME), Terrestrial Ecosystem Science (TES), Subsurface Biogeochemical Research (SBR), and Regional and Global Climate Modeling (RGCM)], (2) programs [Atmospheric Radiation Measurement (ARM)], and (3) projects [Small Worlds; Interoperable Design of Extreme-scale Application Software (IDEAS)]. These research activities generate datasets necessary for Earth system model (ESM) applications and contribute to ESM development by providing model parameterization and benchmarking as well as observation-model software engineering.

Computational Resources

ANL has a variety of projects developing new techniques and tools for storing, transferring, accessing, visualizing, and analyzing large datasets. ANL focuses on producing (1) effective and scalable analysis algorithms; (2) an environment for sharing and interacting with data in complex ways; (3) a robust, distributed software infrastructure; and (4) shared data facilities that provide the "abstract system" on which each operates. Some examples of ANL data science projects are **Glean** (*in situ* visualization and analysis), **Swift** (parallel scripting language for data-intensive science), **TAO** (Toolkit for Advanced Optimization, scalable optimization algorithms that provide the foundation for higher-level data analysis), and **MG-RAST** (metagenomics analysis server). More information on ANL data science projects is available at http://www.mcs.anl.gov/group/data-intensive-science/.

ANL's world-leading computational resources are used in the multilaboratory **ACME** SFA. **PETSc** (Portable, Extensible Toolkit for Scientific Computation) is award-winning numerical software for solving partial differential equations and is the solver in the massively parallel reactive flow and transport model for describing surface and subsurface processes (PFLOTRAN). ANL develops the reference implementation of the **Message Passing Interface** used in all parallel applications. The ACME coupler is built on ANL's **Model Coupling Toolkit**, which is the foundation of ACME, the Community Earth System Model (CESM), and several European coupled models. **Globus Online** provides robust transfer of "big data" for ACME's data management workflow. The **Parallel NetCDF** library is the primary means for parallel file input/output (I/O) in ACME.

The **ARM** team at ANL develops methods, software, and models that bridge the gap between atmospheric observations and their representations in models. ANL conducts development, implementation, and validation of retrievals for passive and active remote-sensing atmospheric instruments, including products derived from radars, microwave radars, radar wind profilers, and micropulse lidar. ANL develops software that is used for extracting geophysical variables from remote-sensing instruments [Python ARM Radar Toolkit (PyART)], as well as methods directed at assimilating observations of aerosols and trace gases into atmospheric models. These

methods include approaches based on the Adjusted Ensemble Kalman Filter and adjoints using the automatic differentiation of Fortran (AdiFOR) software developed at ANL. The national laboratory develops new approaches for assessing data uncertainties that relate to interpolation or gridding of observational datasets to model domains. ANL focuses on combining measurements from multiple instruments to constrain atmospheric processes and derive model parameters. Its experience enables scaling of regional climate models [e.g., Weather Research and Forecasting Model (WRF)] to high-performance computing systems at DOE Leadership Computing Facilities [e.g., ALCF and the National Energy Research Scientific Computing Center (NERSC)].

ANL operates **AmeriFlux** sites (US-IB1 and US-IB2) with over 10 years of data and contributes to the AmeriFlux and FLUXNET networks by participating in site, regional, and global data synthesis and modeling. ANL provides the scientific community with water, energy, and carbon flux measurements on grassland and agricultural ecosystems. The ANL **TES SFA** (http://tessfa.bio.anl.gov/) conducts fundamental research to quantify and characterize carbon stored in soils, evaluate its potential responses to environmental change, and improve the representation of terrestrial ecosystem processes in ESMs. Currently, the Argonne TES SFA is creating a variety of georeferenced datasets to quantify carbon stored in permafrost-region soils, determine its spatial and vertical distributions, and assess the susceptibility of this carbon to decomposition and release to the atmosphere. The ANL **SBR SFA** (http://www.bio.anl.gov/environmental_biology/subsurface_science/doe_ober_sfa.html) provides new mechanistic knowledge of model-relevant biogeochemical processes from the molecular to core scales that helps inform and parameterize multiscale models and helps ensure that the necessary complexities of hydrobiogeochemical processes are included in future ESM development. This work is accomplished by integrating reactive transport modeling approaches with synchrotron-based approaches to characterize experimental flow-through column systems. With laboratory support, ANL also uses data assimilation and optimization approaches for model parameterization and development.

Computational Plans

ANL is developing new algorithmic approaches for analyzing scientific data as well as new methods for mapping existing approaches onto emerging hardware architectures. These algorithms can be used to enhance the environments in which scientists interact with their data through the identification of features of interest or by accelerating analysis to interactive rates. The ANL MCS team also is researching software-defined storage approaches that can tailor data management services to meet the specific needs of BER scientists. This approach will enable deployment of elastic data services on upcoming high-performance computing (HPC) systems such as ANL's Theta and Aurora platforms. At the same time, ANL is assessing the utility of cloud software technologies in the context of BER applications.

In ACME, ANL is developing the crop component of the ACME Land Model (ALM) to better understand and predict the responses of managed land under future climate change. ANL is improving ALM representations of major crops such as maize, soybean, and wheat by using observations from AmeriFlux sites. These efforts will inform models about the impact of climate on crops and the feedbacks crops have on climate through modification of surface radiation and biogeochemistry. ANL also co-leads development of the software engineering processes for ACME and leads development of ACME's I/O system, coupler, and main driver program with a focus on preparing them for exascale operations. ANL secured early access to next-generation supercomputers at NERSC (Cori) and the Oak Ridge Leadership Computing Facility (Summit) and has a proposal pending at ALCF (Theta). ACME already has allocations on current leadership computing facilities (Titan and Mira).

In the TES SFA, ANL is determining whether coupling of fine-scale observational data with landscape and microtopographic features generated from high-resolution lidar data can be used to reduce spatial uncertainty and improve regional estimates of soil carbon stocks. Similarly, in the multilaboratory RGCM SFA, ANL is investigating the impact of spatial scaling on environmental controls, spatial structure, and statistical properties of soil carbon stocks. Under the RGCM SFA, ANL is developing scaling functions for up- or down-scaling of the environmental controls and spatial heterogeneity of soil carbon stocks for coarse- to fine-scale predictions. The TES SFA is building a Fourier transform infrared spectroscopy (FTIR) database for spectral libraries of permafrost-region soils. The FTIR database will aid development of calibrations for estimating soil organic matter characteristics and other soil properties and enable assessment of the potential relationships between FTIR spectra and environmental factors that affect soils. Since these environmental factors can be mapped, their relationships with FTIR spectra—in combination with geospatial analysis and modeling—will enable regional extrapolation and prediction of soil organic matter composition. From both SFAs, TES and

RGCM, ANL expects to generate spatially explicit information contributing to the geospatial cyberinfrastructure described in the following paragraph.

ANL is developing a geospatial cyberinfrastructure to conduct high-resolution geospatial operations at regional and global scales. Efforts to conduct high-resolution geospatial operations will require HPC resources to manipulate, integrate, and analyze large geospatial datasets, particularly as more high-resolution geospatial data products become available. The geospatial cyberinfrastructure will be developed in close communication with the IDEAS project and ANL software engineers working on that project. This geospatial cyberinfrastructure will use data generated by various efforts (SFAs and laboratory-supported projects) to enable model benchmarking studies and identify areas of high model uncertainty. The spatially explicit data products also could be integrated with data assimilation and optimization techniques to advance model parameterization and validation.

The Argonne Small Worlds project is developing a new multimodal imaging capability that incorporates visible light, electron, and X-ray probes for studying intercellular and intracellular dynamical rhizosphere processes. The computational aspects include modeling the design of the imaging systems and optimizing each imaging modality to extract maximal information. Advances in computational imaging that will enhance future ESMs include (1) new algorithms for optimal reconstruction of three-dimensional (3D) structure from sparse data; (2) computational support for bridging scales, primarily in multimodal volumetric registration; (3) methods of inference from comparative analyses of multimodal static and dynamic imagery; and (4) analytical methods for 3D dynamics of molecular-scale objects.

ANL also is developing process-based ecosystem models to represent the mechanistic interactions among terrestrial biogeochemical, biophysical, ecological, and hydrological processes to simulate the responses of high-latitude and tropical ecosystems to climate change at different temporal and spatial scales. ANL plans to use models to guide future studies and to answer other fundamental scientific questions such as microbial dynamics and soil-plant-microbe interactions as well as fate and transport of organic carbon, nutrients, and contaminants. ANL is developing methods for integrating microbial sequencing data (e.g., data from the Earth Microbiome Project) and metabolic predictive models into ecosystem models to mechanistically represent soil biogeochemical processes.

Websites for ANL Computational Resources and Data Science Projects

- ALCF (http://www.alcf.anl.gov)
- Glean (http://www.alcf.anl.gov/glean/)
- Globus Online (https://www.globus.org/)
- Message Passing Interface (http://www.mpich.org)
- MG-RAST (http://metagenomics.anl.gov)
- Model Coupling Toolkit (http://www.mcs.anl.gov/research/projects/mct/)
- Parallel NetCDF (https://trac.mcs.anl.gov/projects/parallel-netcdf/)
- PETSc (http://www.mcs.anl.gov/petsc/)
- Swift (http://www.alcf.anl.gov/swift/)
- TAO (http://www.mcs.anl.gov/research/projects/tao/)

D.2 Lawrence Berkeley National Laboratory

Contributors: C. Steefel (*CISteefel@lbl.gov*), W. Riley (*WJRiley@lbl.gov*), and C. Koven (*CDKoven@lbl.gov*)

Lawrence Berkeley National Laboratory's (LBNL) expertise and computational plans span the range from pore to plume to watershed scales in support of BER's program vision for a "virtual laboratory" devoted to terrestrial ecosystem modeling. The LBNL portfolio currently consists of a mix of in-house simulators and an expanding list of community-supported, open-source software. While the historical strength of LBNL's capability has been in subsurface plume-scale simulation, new directions include the coupling of flow, biogeochemistry, and microbial community composition and function at the pore scale, as well as coupling of subsurface and surface water and vegetation at the watershed scale. Newer-generation models are both multiphysics and multiscale and are designed for present- and next-generation HPC machines.

Computational Resources

LBNL's existing modeling capabilities and expertise support numerous BER programs and applications in terrestrial ecosystems extending from pore to plume to watershed and even to the ESM scale. A key feature of all LBNL efforts, present and future, is the rigorous treatment of biogeochemical, microbial, and vegetation processes in the context of the terrestrial water cycle. Subsurface flow and transport historically have been LBNL's main strength, but coupling to surface water and vegetation

represents the primary focus at present and in the near future. The software generally has been written as proprietary or partly proprietary codes in the Fortran language, with only limited ability to run efficiently on HPC machines. These codes are described briefly in the following paragraph, because they are still used for some terrestrial modeling applications or components of them may be incorporated into newer-application software. The bulk of the current effort in code development is focused on community-based, open-source software specifically designed for HPC machines, which also is briefly described.

Legacy Codes

TOUGH2 is a general-purpose multiphase flow simulator that relies on a suite of equation of state (EOS) modules to compute fluid-phase partitioning and flow and transport of various components (e.g., water, carbon dioxide, salt, air, tracers, and radionuclides) in liquid, gas, and nonaqueous phases (Finsterle, Sonnenthal, and Spycher 2014). By building on **TOUGH2**, **TOUGHREACT** simulates nonisothermal, multicomponent reactive transport of aqueous and gaseous components in variably saturated media. Reactive transport is solved by an operator-splitting approach that can be either iterative or noniterative. The precipitation and dissolution of minerals is optionally coupled to porosity, permeability, and capillary pressure using various correlations that feed back into the multiphase flow computations. Aqueous and gaseous components are transported via advection and diffusion, following the calculation of multiphase fluxes. **CrunchFlow** is a reactive transport software package based on a finite volume discretization of the governing coupled partial differential equations that link flow, solute transport, and multicomponent equilibrium and kinetic reactions in porous and fluid media (Steefel et al. 2015). Two approaches for coupling biogeochemical reactions and transport are available at runtime: (1) a global implicit approach that solves transport and reactions simultaneously (up to 2D) and (2) a time- or operator-splitting approach based on the sequential noniterative algorithm (SNIA; up to 3D). Multicomponent diffusion can be modeled with the Nernst-Planck equation, enabling the inclusion of differing diffusion coefficients for charged species while maintaining electroneutrality (Giambalvo et al. 2002). **CrunchFlow** also can simulate accumulation and transport of ions within a discrete electrical double layer (EDL), with dynamic balancing of the surface charge on the mineral and in the Stern layer calculated with a surface complexation model (Tournassat and Steefel 2015).

Actively Developed Codes

Pore-scale: Pore-scale simulators that make use of soil and aquifer pore structure captured from microtomographic characterization are being actively developed and applied (Molins et al. 2014). The Chombo-Crunch simulator is based on direct numerical simulation of Navier-Stokes flow, transport, and reaction in complex pore structures, capturing interfaces with an embedded boundary approach. The software has been applied to pore-scale reactive transport problems with as many as 2 billion degrees of freedom and good weak scaling up to 10,000 processors.

Plume-scale: The software applicable to plume-scale flow and biogeochemical processes now under active development includes **Amanzi** (Bea et al. 2013), **Parflow-Crunch** (Beisman et al. 2015), and **PFLOTRAN** (Bisht and Riley 2015).

Watershed-scale: The suite of plume-scale codes is being adapted to simulate watershed-scale processes, with a preliminary focus on capturing the flow and partitioning of water between surface, subsurface, and vegetation compartments. The watershed models include **Parflow–Community Land Model** (CLM), **Process-based Adaptive Watershed Simulator (PAWS)–CLM** (Riley and Shen 2014; Shen, Ji, and Riley 2015; Shen et al. 2015), and **PFLOTRAN-CLM** (Pau, Bisht, and Riley 2014; Bisht and Riley 2015). These models are being applied in high-latitude systems [Next-Generation Ecosystem Experiments (NGEE)–Arctic], mid-latitude high-elevation systems (East River, Gunnison County, Colorado), and tropical systems (NGEE–Tropics).

ALM-scale: Regarding CLM and ALM, LBNL made substantial contributions to the most recent version of CLM (i.e., CLM4.5), including a more realistic subsurface biogeochemical representation (Koven et al. 2013); vertically resolved reactive transport solver (Tang and Riley 2013a); soil methane cycle (Riley et al. 2011); improved lake hydrology and sediment biogeochemical module (Subin, Riley, and Mironov 2012); improvements to the soil surface energy budget calculations (Tang and Riley 2013b; 2013c); method to represent multisubstrate and multiconsumer networks applicable to microbe-plant competition for nutrients, microbial competition for carbon, and microbial community diversity (Tang and Riley 2013a); representation of hydraulic redistribution (Tang, Riley, and Niu 2015); and more realistic representation of nutrient controls on soil and plant dynamics (Ghimire, Riley, and Koven 2015; Ghimire et al. 2015; Zhu and Riley 2015; Zhu et al. 2015).

Computational Plans

LBNL's primary focus now is on developing flexible, interoperable multiphysics and multiscale codes for terrestrial modeling. The multiscale nature of the challenge is driving development of new approaches to water and biogeochemical cycling in larger systems that capture hot moments and hot spots, without undue sacrifice of modeling fidelity and resolution. One important effort under way in the IDEAS project is incorporating adaptive mesh refinement (AMR) techniques based on the **Chombo** software into the **Parflow** and **PFLOTRAN** watershed models. LBNL's fine-scale modeling capabilities, as represented in the legacy codes, will provide components for the new flexible, interoperable frameworks, enabling their use in such integrated modeling and experimental (MODEX) efforts as NGEE–Arctic, NGEE–Tropics, and the upper Colorado River modeling initiative under way in LBNL's Genomes to Watershed SFA and IDEAS project.

References

Bea, S. A., et al. 2013. "Identifying Key Controls on the Behavior of an Acidic-U(VI) Plume in the Savannah River Site Using Reactive Transport Modeling," *Journal of Contaminant Hydrology* **151**, 34–54. DOI: 10.1016/j.jconhyd.2013.04.005.

Beisman, J., et al. 2015. "ParCrunchFlow: An Efficient, Parallel Reactive Transport Simulation Tool for Chemically and Physically Heterogeneous Saturated Subsurface Environments," *Computational Geosciences* **19**, 403–22. DOI: 10.1007/s10596-015-9475-x.

Bisht, G., and W. J. Riley. 2015. "Topographic Controls on Soil Moisture Scaling Properties in Polygonal Ground," in review *Hydrology and Earth System Science*.

Finsterle, S., E. L. Sonnenthal, and N. Spycher. 2014. "Advances in Subsurface Modeling Using the TOUGH Suite of Simulators," *Computers in Geoscience* **65**, 2–12. DOI: 10.1016/j.cageo.2013.06.009.

Ghimire, B., W. J. Riley, and C. D. Koven. 2015. "Representing Leaf and Root Physiology in CLM Results in Improved Global Carbon and Nitrogen Cycling Predictions," submitted *Biogeosciences*.

Ghimire, B., et al. 2015. "Global Leaf Nitrogen Allocation for Integration with Terrestrial Land Models: A Synthesis from the Global Plant Traits (TRY) Database," submitted *New Phytologist*.

Giambalvo, E. R., et al. 2002. "Effect of Fluid-Sediment Reaction on Hydrothermal Fluxes of Major Elements, Eastern Flank of the Juan de Fuca Ridge," *Geochimica et. Cosmochimica Acta* **66**(10), 1739–57. DOI: 10.1016/S0016-7037(01)00878-X.

Koven, C. D., et al. 2013. "The Effect of Vertically-Resolved Soil Biogeochemistry and Alternate Soil C and N Models on C Dynamics of CLM4," *Biogeosciences* **10**(4), 7109–31. DOI: 10.5194/bg-10-7109-2013.

Liu, Y., et al. 2015. "A Hybrid Reduced-Order Model of Fine-Resolution Hydrologic Simulations at NGEE–Arctic Study Sites," submitted *Hydrology and Earth System Science*.

Molins, S., et al. 2014. "Pore-Scale Controls on Calcite Dissolution Rates from Flow-Through Laboratory and Numerical Experiments," *Environmental Science and Technology* **48**(13), 7453–60. DOI: 10.1021/es5013438.

Pau, G. S. H., G. Bisht, and W. J. Riley. 2014. "A Reduced-Order Modeling Approach to Represent Subgrid-Scale Hydrological Dynamics for Land-Surface Simulations: Application in a Polygonal Tundra Landscape," *Geoscientific Model Development* **7**, 2091–2105. DOI: 10.5194/gmd-7-2091-2014.

Riley, W. J., et al. 2011. "Barriers to Predicting Changes in Global Terrestrial Methane Fluxes: Analyses Using CLM4Me, a Methane Biogeochemistry Model Integrated in CESM," *Biogeosciences* **8**, 1925–53. DOI: 10.5194/bg-8-1925-2011.

Riley, W. J., and C. Shen. 2014. "Characterizing Coarse-Resolution Watershed Soil Moisture Heterogeneity Using Fine-Scale Simulations," *Hydrology and Earth System Science* **18**, 2463–83. DOI: 10.5194/hess-18-2463-2014.

Shen, C., X. Ji, and W. J. Riley. 2015. "Temporal Evolution of Soil Moisture Statistical Fractal and Controls by Soil Texture and Regional Groundwater Flow," to be submitted *Hydrology and Earth System Science.*

Shen, C., et al. 2015. "The Fan of Influence of Streams and the Impacts of Channel Density on Simulated Water and Carbon Fluxes," in review *Water Resources Research*.

Steefel, C. I., et al. 2015. "Reactive Transport Codes for Subsurface Environmental Simulation," *Computational Geosciences* **19**(3), 445–78. DOI: 10.1007/s10596-014-9443-x.

Subin, Z. M., W. J. Riley, and D. Mironov. 2012. "An Improved Lake Model for Climate Simulations: Model Structure, Evaluation, and Sensitivity Analyses in CESM1," *Journal of Advances in Modeling Earth Systems* **4**, M02001. DOI: 10.1029/2011MS000072.

Tang, J. Y., and W. J. Riley. 2013a. "A Total Quasi-Steady-State Formulation of Substrate Uptake Kinetics in Complex Networks and an Example Application to Microbial Litter Decomposition," *Biogeosciences* **10**(12), 8329–51. DOI: 10.5194/bg-10-8329-2013.

Tang, J. Y., and W. J. Riley. 2013b. "Impacts of a New Bare-Soil Evaporation Formulation on Site, Regional, and Global Surface Energy and Water Budgets in CLM4, *Journal of Advances in Modeling Earth Systems* **5**(3), 558–71. DOI: 10.1002/jame.20034.

Tang, J. Y., and W. J. Riley. 2013c. "A New Top Boundary Condition for Modeling Surface Diffusive Exchange of a Generic Volatile Tracer: Theoretical Analysis and Application to Soil Evaporation," *Hydrology and Earth System Sciences* **17**, 873–93. DOI: 10.5194j/hess-17-873-2013.

Tang, J. Y., et al. 2013., "CLM4-BeTR, a Generic Biogeochemical Transport and Reaction Module for CLM4: Model Development, Evaluation, and Application," *Geoscientific Model Development* **6**, 127–40. DOI: 10.5194/gmd-6-127-2013.

Tang, J. Y., W. J. Riley, and J. Niu. 2015. "Implementing Root Hydraulic Redistribution in CLM4.5: Model Development, Testing, and Application, in review *Geophysical Model Development*.

Tournassat, C., and C. I. Steefel. 2015. "Ionic Transport in Nano-Porous Clays with Consideration of Electrostatic Effects," *Reviews in Mineralogy and Geochemistry* **80**, 287–329. DOI: 10.2138/rmg.2015.80.09.

Trebotich, D. P., et al. 2014. "High-Resolution Simulation of Pore-Scale Reactive Transport Processes Associated with Carbon Sequestration," *Computing in Science and Engineering* **16**, 22. DOI: 10.1109/MCSE.2014.77.

Zhu, Q., and W. J. Riley. 2015. "Using ECA Kinetics and an Improved Leaching Model in CLM4.5 Improves Comparisons to Observations: Response to 'Improving Nitrogen in Climate Change Forecasts'," submitted *Nature Climate Change*.

Zhu, Q., et al. 2015. "Multiple Soil Nutrient Competition Between Plants, Microbes, and Mineral Surfaces: Model Development, Parameterization, and Example Applications in Several Tropical Forests," *Biogeosciences Discussion* **12**, 4057–4106. DOI: 10.5194/bgd-12-4057-2015.

D.3 Los Alamos National Laboratory

Contributors: E. Coon *(ecoon@lanl.gov)*, D. Moulton *(moulton@lanl.gov)*, J. Ahrens, P. Jones, S. Karra, R. Linn, P. McCormick, N. McDowell, T. Ringler, J. Rowland, V. Vesselinov, C. Wilson, and C. Xu

The expertise and computational plans of Los Alamos National Laboratory (LANL) are well aligned with BER's virtual laboratory vision (BERAC 2013) and phased approach to a community-supported software ecosystem (U.S. DOE 2015). This alignment is demonstrated through high levels of software engineering in ongoing development of process-based models, multiscale and multiphysics modeling techniques, new methods for uncertainty quantification and decision support, and advanced numerical algorithms on emerging hardware architectures.

Computational Resources

LANL's existing capabilities and expertise support various BER applications ranging from fine- to global-scale models and are split between methodological tools and application software and expertise. Methodological capabilities include software frameworks for model component integration in multiphysics and multiscale problems on HPC machines. Many of these capabilities build on strong expertise developed at LANL via support from DOE's programs in the Office of Science (SC) and Office of Advanced Scientific Computing Research (ASCR) and the National Nuclear Security Administration's (NNSA) Advanced Simulation and Computing (ASC). This development and collaboration are ongoing and central to meeting the demands of new programming models on emerging architectures. Examples include LANL's next-generation multiphysics framework **Arcos** (NGEE–Arctic, NGEE–Tropics, and IDEAS), which manages complexity for both process-rich and process-uncertain models, and the Model for Prediction Across Scales (**MPAS**), which enables flexible, multiscale simulation on global domains using appropriate model components at appropriate scales. **Legion** is a data-centric parallel programming tool for expressing task-level concurrency. **Paraview** and **Cinema** enable visualization and *in situ* analysis of running simulations; these capabilities will prove critical as storage and I/O become performance bottlenecks. Finally, LANL has leading expertise and software frameworks for data, model, and decision analyses, including a model analysis toolkit (**MATK**, NGEE–Arctic); **DiaMonD**, an ASCR center on integrated approaches to novel theoretical methods at the interfaces between data, models, and decisions; a model analysis and decision support (**MADS**) HPC framework; and a **UASA ToolBox** for uncertainty and sensitivity analysis.

Application software and expertise with LANL leadership and significant development efforts include ocean and ice **ACME** models, as well as terrestrial ecosystem demography, plant mortality, hydrology, and landscape disturbance and evolution submodels for ALM- and fine-scale models. A strong LANL focus is the science of interfaces between land-atmosphere, land-aquatic, and land-atmosphere-aquatic systems in the context of both long-term mean climate change (e.g., sea-level rise, atmospheric warming, and drying) and extreme events (e.g., droughts, fires, floods, and hurricanes). Specifically, LANL contributes to a multilaboratory, community-driven suite of fine- to basin-scale hydrology models, including **Amanzi** [Advanced Simulation Capability for Environmental Management (ASCEM)], the Advanced Terrestrial Simulator (**ATS**, NGEE–Arctic and NGEE–Tropics), and **PFLOTRAN** (NGEE–Arctic); these are key codes in the IDEAS effort to develop componentized software. **Higrad**, **FIRETEC**, and **QUICFIRE** predict exchanges between heterogeneous vegetation and local atmosphere at submeter resolution, fire phenomena, and efficient stochastic fire impacts, respectively.

Plant mortality and ecosystem demography are key processes that are not well (or at all) represented in current ESMs; LANL researchers lead in modeling these phenomena in **SUMO** (SUrvival MOrtality experiments). Additional application expertise focuses on land interfaces with climate components; LANL has expertise in ocean-land interactions, including geomorphic events (landslides, thermokarst modeling, and delta dynamics) and sea-level rise impacts leveraging Climate, Ocean and Sea Ice Modeling (**COSIM**).

Computational Plans

These capabilities, with their flexible composition, identify areas in which LANL's expertise is critical to ESS program goals (U.S. DOE 2012). For example, predicting extreme event impacts and their climate feedbacks is both critically important for climate predictions and difficult for ESMs to represent. LANL's expertise in developing and applying multiscale and multiphysics frameworks that couple fire, thermal hydrology, ecosystem demography and mortality, and geomorphic response enables improved predictions of terrestrial-climate system interactions. These frameworks now are being applied to understand and predict coupled land-aquatic system evolution with changing climate. LANL's fine-scale model capabilities and expertise in coupling these components are well positioned to enable MODEX efforts such as NGEE–Arctic and NGEE–Tropics, where models act as integrating tools to turn localized observations into improved representations of key processes in ESMs. Additionally, these same fine-scale capabilities enable predictions of climate impacts on regional and local systems, including critical watersheds for energy use and water resources.

Underlying these plans for continued work in ESS applications are plans for continued work in infrastructure and frameworks enabling more efficient model development, coupling of existing components across scales and traditional disciplines, improved uncertainty quantification and decision support, and portable performance on next-generation hardware. Multiphysics frameworks such as Arcos and multiscale frameworks such as MPAS enable existing components to be rapidly coupled in flexible ways, leveraging ongoing capability development for many application programs. LANL will leverage these and other capabilities to provide a powerful way to express multiscale, multiphysics models in a form amenable for high-performance frameworks such as Legion to execute task-parallel simulations on next-generation machine architectures, providing a path to exoscale computation that is both viable and tractable for the scientific community.

Finally, LANL supports the fundamental shift from building monolithic application codes to building models from interoperable components within a framework that provides access to significant resources through libraries and tools. This new approach to model development, as well as corresponding improvements in model-data integration workflows, is essential to enhance scientific productivity and interdisciplinary collaboration across BER applications during this time of disruptive changes to both hardware and software.

References

BERAC. 2013. *BER Virtual Laboratory: Innovative Framework for Biological and Environmental Grand Challenges:A Report from the Biological and Environmental Research Advisory Committee*, DOE/SC 0156. Biological and Environmental Research Advisory Committee, U.S. Department of Energy Office of Science, Washington, D.C. (science.energy.gov/ber/berac/reports/).

U.S. DOE. 2015. *Building Virtual Ecosystems: Computational Challenges for Mechanistic Modeling of Terrestrial Environments: Workshop Report,* DOE/SC-0171. U.S. Department of Energy Office of Science (www.doesbr.org/BuildingVirtualEcosystems/).

U.S. DOE. 2012. *Climate and Environmental Sciences Division; Strategic Plan*, DOE/SC-0151. U.S. Department of Energy Office of Science, Office of Biological and Environmental Research, Washington, D.C. (science.energy.gov/~/media/ber/pdf/CESD-StratPlan-2012.pdf).

D.4 Oak Ridge National Laboratory

Contributors: D. Bernholdt, T. Boden, J. Fellows, J. Gulledge, F. Hoffman, L. Hook, G. Jacobs, J. Kumar, M. Mayes, S. Painter, B. Preston, D. Ricciuto, P. Thornton, D. Wang, and S. Wullschleger

Oak Ridge National Laboratory (ORNL) presents (1) a use case that exemplifies the challenges of integrating scientific discovery data, synthesized data for model-experiment integration, simulation output, and model development and (2) examples of systems and actions that offer partial solutions to the challenges.

Use Case: Challenges in Hydrobiogeochemical Process Studies and Model Development at Multiple Scales

The NGEE–Arctic project engages observation, experimentation, and modeling for multiple coupled processes across multiple overlapping spatial and temporal scales. Each investigator may contribute to several aspects of the project,

and each aspect of the overall science plan may draw from the efforts of many individuals, integrating across scales of study, disciplines, and institutions. A top-level NGEE–Arctic goal is to improve global-scale prediction of climate-biogeochemistry feedbacks by integrating process knowledge to the scale of a climate-model gridcell. Meeting this goal demands an organizational framework that enables the query of existing modeling, observational, and experimental results along multiple dimensions (1) to relate measurements and models through hypothesis generation, boundary condition specification, parameter estimation (PE), and model evaluation and (2) to make quantitative estimates of prediction uncertainty. These demands for an organizational framework can perhaps be clarified by restating the challenges as a series of questions that an NGEE–Arctic investigator might ask who is tasked with high-level knowledge integration:

- What simulations have already been done for the region of interest?
- At what resolutions? In what model configurations (e.g., offline versus coupled)? Over what time periods (hind-cast, future scenarios)? With what level of mechanistic detail?
- What is known about the uncertainty in existing simulation results (e.g., uncertainty in forcings, parameters, model structure)?
- What testable hypotheses have been framed by previous modeling?
- What observational data exist for the region and processes of interest?
- At what spatial scale and over what time period?
- Which measurements are explicitly coordinated in space and time (e.g., multiple measurements from the same plots or quantities measured from the same samples)?
- What level of synthesis has already been applied to existing observational data? Is there any quantification of observational uncertainty?
- What process-resolving experiments have been carried out?
- At what spatial and temporal scales?
- What are the significant experimental findings?
- Do existing results from modeling, observation, and experimentation point to inconsistencies or arrive at contradictory conclusions?
- Does existing uncertainty quantification (UQ) help to constrain the interpretation of any inconsistencies?
- If significant inconsistencies are absent (for a particular process domain), have the relevant pieces of evidence been presented as an assessment or synthesis? If not, this could be a good opportunity for a paper that states some level of confidence in current knowledge of the system.
- Do modeling and data synthesis point to new, testable hypotheses?
- Which new observations and experimentation are needed to investigate such hypotheses?
- Are measurement and experimentation teams in place to investigate such hypotheses?
- For new observations, experimentation, and modeling under way, what are the results showing?

Examples of Systems and Actions that Can Provide Partial Solutions

Integration, Guidance, and Governance. Integration requires collaboration. Data sharing, communication, and proper acknowledgement are requisite to successful collaboration. Consistency and standardization help support model-data integration. DOE should support and expect the collaborative implementation of consistent and standardized processes for collecting and tracking samples, reporting data and metadata, quality assessment, documentation, product generation, and model-data integration across BER projects.

Data Lifecycle (e.g., Data Planning, Quality Assurance/Quality Control, Provenance, and Interaction with Modelers). One of the challenges in achieving model-data integration is planning for the full data and metadata lifecycle before sampling strategies and measurements begin. Clear communication among science teams is critical for defining model-data needs, scientific objectives and tasks, possible measurements, and data delivery systems. Once defined, data needs should be met in the most efficient and standardized manner possible. Data from existing sources should be obtained, quality checked, documented, and provided through a common data portal or framework. It is imperative that modelers define and communicate model features (e.g., regional or global scale) and requirements (e.g., temporal resolution and needed measures of uncertainty) and the full suite of model data needs for parameterization, initialization, inputs and drivers, calibration, and testing. Plans for (1) new sampling, analyses, and data collection; (2) data and metadata reporting; and (3) processing and quality assessment of data products should be designed to meet both modeling and science task needs. A use case

focusing on the delivery of specific, fully qualified NGEE data in the form required to satisfy a suite of models would help to illuminate challenges and possible solutions.

Model Development and Implementation. Model development and implementation challenges associated with ORNL's ESS research arise from two simultaneous demands: (1) the need for highly resolved, process-rich simulations to facilitate improved process understanding and (2) the need for representations of Earth surface processes that are computationally tractable within global ESMs. These two demands are competing—*small-scale* modeling greatly benefits from model flexibility, the ability to incorporate observations, and the ease of linking to UQ and PE tools; *global-scale* modeling places a higher priority on computational efficiency and robustness. Imagine, as a long-term goal, a software-data "ecosystem," where domain scientists are able to focus on process understanding and model validation and then deploy relatively easily those refined submodels in global models. In this software-data ecosystem, domain scientists would write fewer lines of new code by reusing and repurposing existing process submodels, taking advantage of common interfaces to supporting computational and system function (e.g., solvers, meshes, discretization algorithms, I/O, and error handling), data integration tools (PE and UQ frameworks), and easier access to observational data. A use case focusing on integration of models with experiments and observations (MODEX) in the Arctic context could make important advances toward this vision. Specifically, that use case could help drive the development of software designs and refactoring tools that facilitate transfer of model components between large- and small-scale modeling tools. Further, such a use case would produce prototypical workflows for UQ and PE, which then could be tested in real-world settings. Moreover, the exercise of using submodels with PE and UQ frameworks would build important community experience with designing and developing interface-aware submodels.

Tools for Large-Scale Environmental Data Analytics. Observational and modeled data acquired or generated by the various Earth science disciplines encompass temporal scales of seconds to millions of years (10^0 s to 10^{13} s) and spatial scales of microns to tens of thousands of kilometers (10^{-6} m to 10^7 m). Because of rapid technological advances in sensor development, computational capacity, and data storage density, the volume, complexity, and resolution of Earth science data are increasing equally rapidly. Moreover, combining, integrating, and synthesizing data across Earth science disciplines offer new opportunities for scientific discovery that are only beginning to be realized. Data-centric science, however, also poses unique technological and social challenges, many of which are exacerbated by the sheer size of the datasets involved. A wide variety of data mining, machine learning, and information theoretic techniques now are being applied to a growing body of Earth science data. Cluster analysis has proven useful for segmentation, feature extraction, network analysis, change detection, model intercomparison, and model-data comparison in a number of Earth science applications. Block entropy can be used as a classifier for dynamical systems. Spectral methods are frequently employed for decomposing periodic phenomena. Artificial neural networks and model tree ensembles have been used to refine models and to empirically up-scale and extrapolate point measurements.

D.5 Pacific Northwest National Laboratory

Contributors: K. Kleese van Dam, T. Scheibe, X. Chen, R. Leung, M. Huang, C. Sivaraman, V. Bailey, G. Asrar, J. Comstock, T. Seiple, M. Corsello, and L. Riihimaki

Pacific Northwest National Laboratory (PNNL) has a broad portfolio of activities in data, data management and analysis, model-data integration, and modeling related to BER ESS program priorities. The following sections highlight only a few key PNNL focus areas relevant to discussions at the April 2015 ESS workshop.

Facilities

The **DOE Environmental Molecular Sciences Laboratory** (EMSL) operates the Cascade supercomputer for EMSL's user program. EMSL is developing a multiscale modeling framework, starting from the premier high-performance computational chemistry code NWChem and building upward to cellular, pore, and ecosystem scales. EMSL's computational capability is closely linked with a suite of state-of-the-art molecular- and pore-scale experimental facilities including a microfluidics fabrication and experimentation facility, an intermediate-scale subsurface flow and transport laboratory, an X-ray microtomography imaging laboratory, and extensive microscopy and spectroscopy facilities.

The **ARM Climate Research Facility**, a multilaboratory research program led by PNNL, was established in 1989 by BER to provide an observational basis for studying Earth's climate. ARM's primary capabilities include a network of long-term, fixed-location observation sites; three mobile

observation facilities that are typically deployed for about a year at a time; and an aerial facility. ARM provides comprehensive measurements at high spatial and temporal resolution and is ideally suited to study atmospheric processes. ARM offers scientists the ability to propose observational campaigns to study specific atmospheric processes including support for NGEE–Tropics. ARM's main focus is on the creation and dissemination of atmospheric datasets that support the improvement of physical process representation in climate models.

Value-Added Datasets

The **ARM** program provides a wide range of data products including weather radar precipitation rate (external ARM product); carbon dioxide fluxes; radiative fluxes (broadband and spectral); large eddy simulation (LES) forcing datasets and simulation output; soil moisture and temperature; surface sensible and latent heat flux; near-surface temperature, humidity, and winds; and boundary-layer depth, surface precipitation rate (tipping bucket and optical rain gauges), and rain rates derived from vertically pointing and scanning ARM radars.

Multisource Datasets for Integrated Research, Modeling, and Analysis include land cover, population, energy, agriculture, and emissions data.

The **Hanford Environmental Systems** (HEIS) database includes a subsurface geologic map built with Earth vision software using well-based geologic logs, contaminant plume simulations, and conservative and reactive tracer tests conducted at the Hanford Integrated Field Research Challenge (IFRC) Site for studying the persistence of contaminant plumes and geophysical monitoring data for monitoring river water and groundwater interactions.

Management of Data and Modeling Processes

ARM Data Integrator (ADI). ADI is an open-source software framework that simplifies the generation of customized datasets (see Fig. D5-1. ARM ADI Workflow and Framework Hooks, this page). ADI offers standard services that automate data retrieval, merging of diverse datasets, regridding, and the creation of data products that conform to ARM standards. Process details are defined and maintained through a web interface and stored in a database. Users can edit these templates to add their own logic and scientific algorithms or call in non-ADI functions written in other languages to manipulate the input data and create new or derived values. Algorithm development is supported in C, Python, and Interactive Data Language (IDL). The ADI libraries are open source and available at https://github.com/ARM-DOE/ADI. The ability to reuse existing services has reduced the costs of creating new custom datasets by 30% to 70%, depending on the complexity of the data retrievals and scientific algorithms to be implemented.

Data Versioning. The ARM archive currently has data holdings exceeding 750 TB in 10 million files and greater than 5,000 data products (data streams). Data from ARM sites are collected and stored daily at a current pace exceeding 17 TB per month. Data

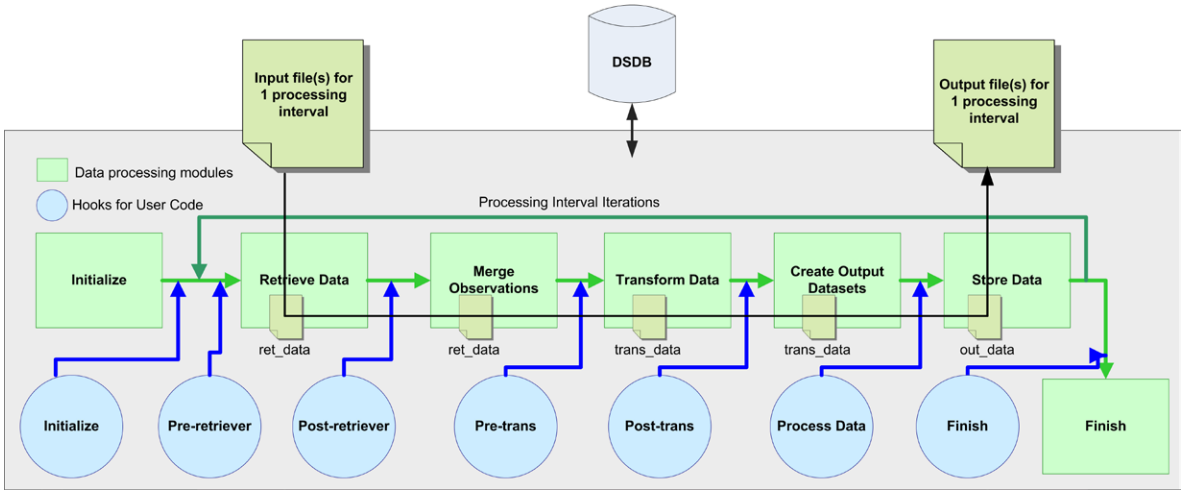


Fig. D5-1. ARM ADI Workflow and Framework Hooks. [Gaustad, K., et al. 2014. "A Scientific Data Processing Framework for Time Series Net CDF Data," *Environmental Modelling and Software* **60**, 241–49. Reprinted by permission of Elsevier Ltd.]

streams are evaluated routinely for quality and considered for improvement through reprocessing. With 20 years of historical datasets, more than 360 instruments, and 3,200 processes, the ability to convey differences in data streams and products over time is vitally important. For this purpose, ARM developed a data versioning system. The system applies version numbers to files and generates new version numbers if the file content changes. Version visualization and discovery tools enable users to quickly understand data changes.

PHOENIX. Visualization of spatial and temporal data is an important step in model-data integration. PNNL Hanford Online Environmental Information eXchange (PHOENIX) is a suite of web-based data access and visualization tools developed for DOE Richland Operations (DOE-RL) and Office of River Protection (DOE-ORP). The tools integrate disparate data from environmental (e.g., groundwater, soil, and weather) and tank monitoring databases into at-a-glance dashboard and map-based views. Displaying data in context allows PHOENIX users to discover new correlations using existing data.

Velo/AKUNA. In the ASCEM project, PNNL developed the Velo-based Akuna collaborative framework. The Akuna-Agni platform provides the user interface and toolsets for managing workflows, including model development starting with definition of the conceptual model, management of data and metadata for model input, sensitivity analysis, model calibration and uncertainty analysis, model execution on diverse computational platforms, and processing of model output, including visualization. As part of the PNNL SBR SFA and in collaboration with LBNL, CLM (the precursor of ALM) was integrated into the Akuna-Agni framework and has been applied to several research sites to demonstrate its capabilities. Completed demonstrations include (1) 1D CLM simulations of the NGEE–Arctic site, with sensitivity analysis performed on several soil parameters (e.g., organic content and percent clay and sand); and (2) 1D CLM simulations of the ARM Southern Great Plains central facility site, with sensitivity analysis of 10 parameters (with results compared to a previous publication using a different method) and application of the PE toolset. PNNL also is working toward demonstrations of 3D CLM simulations at the NGEE–Arctic and Hanford 300 Area sites.

Model-Data Integration

Bayesian Data Assimilation. PNNL has developed various Bayesian data assimilation techniques (e.g., full Bayesian approach and Ensemble Kalman Filter–based methods) to characterize spatially heterogeneous subsurface properties of the Hanford 300 Area by integrating multiscale and multitype laboratory and field experimental data with modeling. The Bayesian approach is flexible in dealing with uncertainties arising from different sources and with data that come in a temporal sequence. This approach also enables direct quantification of uncertainty in model predictions by providing an ensemble of estimated model parameters. PNNL has leveraged HPC resources [e.g., community flow and transport code PFLOTRAN and coupled hydrogeophysics code (PFLOTRAN+E4D)] and HPC facilities at the national laboratories (e.g., Hopper, Edison, and Cascade) to handle the computational challenge. HPC use is required by increasingly mechanistic nonlinear forward simulations and further amplified by the need of stochastic data assimilation and increase in spatial and temporal scales of model simulations. PNNL has successfully estimated a subsurface permeability field with close to a half million unknowns using the Ensemble Kalman Filter–based technique with HPC resources.

Markov Chain Monte Carlo (MCMC) Model Calibration. With ASCR funding and in collaboration with Sandia National Laboratories (Program Manager Steven Lee), PNNL has developed approaches to perform MCMC calibration of the CLM. The first approach is based on surrogates of the CLM—inexpensive polynomial or Gaussian process representations of the mapping between CLM parameters being calibrated and CLM outputs for which there are measurements. PNNL is building on and extending recent developments on the use of surrogates to calibrate computationally expensive models and MCMC calibration of complex models (e.g., those based on partial differential equations), including structural errors (i.e., the model's fundamental inability to reproduce observations, a result of modeling simplifications). The second approach is called Scalable Adaptive Chain-Ensemble Sampling (SAChES), which is a scalable multichain MCMC method that is robust to nondeterministic hardware, by combining differential evolution of Markov chains, adaptive Metropolis, and ensemble sampling. Differential evolution enables the exploration of high-dimensional parameter spaces using loosely coupled (i.e., largely asynchronous) chains so that large-chain ensembles (i.e., far more chains than the number of parameters to explore) are enabled. SAChES already has been coupled with CLM and implemented on HPCs.

Modeling

Pore- to Core-Scale Representation of Carbon Dynamics in Soils. PNNL's TES SFA is developing a pore- to core-scale representation of carbon dynamics in soils, an objective which entails integration of core-scale gas fluxes measured with representative datasets from field collars and flux towers (through collaborations). PNNL also is generating highly resolved soil carbon chemical profiles and enzyme potentials for these soils at the pore and core scales. This will improve representation of "passive" or physically protected carbon in current biogeochemical models, identifying conditions under which a portion of this passive pool may become active. In addition, Ben Bond-Lamberty, a project co-investigator, curates and maintains a global soil respiration database for the research community.

Groundwater–Surface Water Interactions and Their Impact on Subsurface Biogeochemistry. PNNL's SBR SFA is performing integrated research on groundwater–surface water interactions and their impact on subsurface biogeochemical cycling of key nutrients and contaminants. Modeling serves an integrating role in this project, providing multiscale linkages among laboratory-, local field–, and river reach–scale experiments. Advanced modeling capabilities including high-performance codes for pore-scale simulation; field-scale reactive transport modeling (PFLOTRAN and eSTOMP); joint geophysical-hydrologic inversion and Bayesian data assimilation (coupled PFLOTRAN-E4d); and integrated river, groundwater, and land-surface modeling [CLM-PFLOTRAN and MOdel for Scale Adaptive River Transport (MOSART)]. PNNL has pioneered the development of hybrid multiscale simulation methods, in which multiple models at different scales are coupled within a single multiscale simulation.

Land Surface Hydrology. PNNL contributes to ALM development, with a focus on improving the representation of land surface hydrology and impacts of water management. Motivated by previous PNNL research, ALM adopts the use of watersheds as the computational units and further represents topographic variations within watersheds by a limited number of topographic land units defined by surface elevation, slope, and aspect. The PNNL team is developing the methodology and input data to support this new spatial structure. In addition, PNNL is (1) adding a new parameterization of inundation using two-way coupling of ALM and the MOSART river routing model; (2) extending MOSART to represent riverine biogeochemistry and stream temperature; (3) implementing the surface and subsurface runoff parameterizations of the Variable Infiltration Capacity model; and (4) coupling a water management model with ALM and MOSART to fully represent the impacts of reservoir operations on water cycle processes.

Surface and Subsurface Water Availability to Plants. PNNL leads an NGEE–Tropics research objective to improve understanding and modeling of surface and subsurface water available to plants. As part of this effort, the PNNL team leads the design and execution of a set of numerical experiments using a hierarchy of hydrologic models (1D, semi- and fully distributed, and 3D) to evaluate their scalability and process representations. The goal is to improve a modular modeling framework for hydrologic modeling in the next-generation ESM. PNNL also co-leads the Manaus Pilot Study to investigate hydrology-carbon interactions and implications to tropical forest response to droughts. In addition, PNNL also contributes to the NGEE–Tropics research objective in understanding and modeling disturbance and land use change.

Multiscale Coupling. PNNL is a partner in the IDEAS project, which is working on a new extreme-scale scientific software ecosystem in which modern software engineering tools, practices, and processes improve software development productivity, and applications are constructed using components, libraries, and frameworks. PNNL leads the multiscale model framework development activities related to both IDEAS use cases and the model coupling element of the Extreme-Scale Software Development Kit (xSDK).

Appendix E. Computational Trends Informing Environmental System Science Projects and Programs Within BER's Climate and Environmental Sciences Division

Prior to the workshop, each of the participating national laboratories was asked to provide a two-page description of computational trends, developments, challenges, and opportunities in hardware and software that they are tracking and explain which are believed to have potential impact on Environmental System Science (ESS) projects and programs within the Climate and Environmental Sciences Division (CESD) of the Department of Energy's (DOE) Office of Biological and Environmental Research (BER). Although scope was left to the authors, the request suggested that the two-page description should highlight key challenges and opportunities, particularly with respect to the current code base (a mix of established codes ranging in age from 5 to 30 years), spanning monolithic serial codes to distributed parallel codes that leverage libraries and frameworks, and describe how they play into the workflow required for model-data integration.

These descriptions are included in the following sections.

E.1 Oak Ridge National Laboratory

Contributor: D. E. Bernholdt

Oak Ridge Leadership Computing Facility (OLCF)

OLCF (http://olcf.ornl.gov) is one of three computing facilities supported by DOE's Office of Advanced Scientific Computing Research (ASCR). The other two are the Argonne Leadership Computing Facility (ALCF) and the National Energy Research Scientific Computing Center (NERSC) at Lawrence Berkeley National Laboratory. OLCF's primary current resource is Titan, a 27 PF Cray XK7 system, but work already is well under way on Summit, an IBM system with NVIDIA graphics processing units (GPUs) and Mellanox interconnect, which is scheduled to go into production in 2018. See Table 1. Comparison of Summit and Titan Features, this page, for a comparison of key features and specifications of the two systems. The new system will have a great deal in common with Titan, but there also will be differences. OLCF has selected 13 applications for its Center for Accelerated Application Readiness (CAAR) to lead the way in understanding how to make the most effective use of the new system, including Accelerated Climate Modeling for Energy (ACME). Lessons learned from the CAAR applications will aid other projects in porting to Summit in the future.

Worth noting is that, while OLCF is preparing for Summit, the other ASCR facilities also will be getting new systems. NERSC's Cori system will be deployed in two phases, with the second-phase system featuring the Intel second-generation Xeon Phi (manycore) processor to be deployed in mid-2016. ALCF also will deploy a similar system named Theta in 2016, followed by Aurora, with the third-generation Xeon Phi processor expected to go into production in 2019. Performance portability of applications across the manycore and accelerated architectures at the different facilities will be an important consideration, alongside performance itself, and will be emphasized in CAAR and the corresponding early science programs at the other facilities.

Table 1. Comparison of Summit and Titan Features

Feature	Summit	Titan
Application performance	5 to 10× Titan	Baseline
Number of nodes	~3,400	18,688
Node performance	>40 TF	1.4 TF
Memory per node	> 512 GB (HBM+DDR4)	38 GB (GDDR5+DDR3)
NVRAM per node	800 GB	0
Node interconnect	NVLink (5 to12× PCIe 3)	PCIe 2
System interconnect (node injection bandwidth)	Dual Rail EDR-IB (23 GB/s)	Gemini (6.4 GB/s)
Interconnect topology	Nonblocking Fat Tree	3D Torus
Processors	IBM POWER9 NVIDIA Volta™	AMD Opteron™ NVIDIA Kepler™
File system	120 PB, 1 TB/s, GPFS™	32 PB, 1 TB/s, Lustre®
Peak power consumption	10 MW	9 MW

Compute and Data Environment for Science (CADES)

CADES is an integrated compute and data science infrastructure and service portfolio being deployed in support of Oak Ridge National Laboratory (ORNL) projects and staff. It provides a diverse computing and data ecosystem supported by matrixed staff with expertise in various areas of computing and data science. CADES is focused on the technical computing and data needs of the scientific and engineering research and development communities across ORNL. CADES includes the concept of an externally accessible user commons, making its resources and services available outside ORNL to support the project's requirements for external access to collaborators or the public.

SC14 Data Science Demonstrations

ORNL was a strong participant in the Data Science Demonstrations, along with 11 other institutions, presented in the DOE booth at SC14 in New Orleans. These demonstrations (see Table 2. SC14 Data Science Demonstrations, this page), which spanned a broad range of science areas, serve to illustrate the diverse and growing needs of the DOE complex to deal with complex data at large scales and some of the approaches being developed. ASCR is expected to increasingly emphasize data science in the coming years.

Table 2. SC14 Data Science Demonstrations

Demonstration*	Participants†
Bringing the Power of HPC to BES X-Ray Light Source Facilities	Craig E. Tull et al., ANL, BNL, LBNL, ORNL, PNNL, SLAC
BigPanDA: New Advances in Workload Management for Opportunistic Supercomputing	Kenneth Read et al., ORNL
100 G+ Data Transfer via Embedded GridFTP in a DDN Disk Controller	Eunsung Jung, Raj Kettimuthu, ANL
ParaView Scientific Visualization Demo: Running Large Datasets	W. Alan Scott, SNL
Granular Data Processing on HPCs Using an Event Service	Torre Wenaus et al., BNL
Dark Energy Science Analysis Using Diverse Applications Across DOE Computing Facilities	Saba Sehrish, Fermilab
The TAU Performance System	The TAU Team, University of Oregon
EXDAC: EXtreme Data Analysis for Cosmology	Peter Nugent et al., LBNL
ARGO: An Exascale Operating System and Runtime	Pavan Balaji et al., ANL
Streaming of Large-Scale Simulation in Real Time with the ViSUS PIDX Library	Peer-Timo Bremer et al., LLNL, ANL, LANL, SNL
Deep Data Analytics and Scientific Inference Microscopy, BES Science	Kerstin Kleese-Van Dam, PNNL
Real-Time Processing of Human and Rodent Neurological Recording Data	David Donofrio, LBNL
High-Performance Parallel I/O	Prabhat, Berkeley and Quincey Koziol, HDF Group

*BES: Basic Energy Sciences; DDN: Data Direct Networks; HPC: high-performance computing; I/O: input/output; SC14: International Conference for High-Performance Computing, Networking, Storage, and Analysis, annual meeting; TAU: tuning and analysis utilities

†DOE national laboratories: Argonne National Laboratory (ANL), Brookhaven National Laboratory (BNL), Fermi National Accelerator Laboratory (Fermilab), Lawrence Berkeley National Laboratory (LBNL), Lawrence Livermore National Laboratory (LLNL), Los Alamos National Laboratory (LANL), Oak Ridge National Laboratory (ORNL), Pacific Northwest National Laboratory (PNNL), Sandia National Laboratories (SNL), SLAC National Accelerator Laboratory (SLAC)

E.2 Pacific Northwest National Laboratory

Contributors: K. Kleese van Dam, N. Baker, and D. Kerbyson

Pacific Northwest National Laboratory (PNNL) has a broad portfolio of activities related to data science, mathematics, and high-performance computing that can support BER ESS program priorities. This document highlights only a few key PNNL focus areas relevant to discussions at the April 2015 ESS workshop.

In Situ Data Analysis

A signature is a process that transforms data in the form of features into labels with associated classification uncertainties. Signatures are used in a wide range of BER-relevant domains, including the diagnostic assessment of atmospheric monitoring instruments, prognostic assessment of microbial community stability, and forensic characterization of geological processes. The PNNL Signature Discovery Initiative (http://signatures.pnnl.gov) has developed domain-agnostic methodology and software to more robustly and efficiently discover signatures from noisy and incomplete multisource data streams. Applications of Signature Discovery methodology include signatures for identification of cloud-phase states from multiple Atmospheric Radiation Measurement (ARM) remote-sensing observations, signatures of ecosystem resilience based on multisource soil structure and microbial community measurements, and signatures of cellular perturbation and growth history.

Analysis in Motion

Science missions are driven by the need to assimilate and interpret ever-increasing volumes of data to accelerate scientific discovery and make critical decisions, so the speed of analysis is as important as the choice of data to be collected. The Analysis in Motion Initiative (AIM; http://aim.pnnl.gov) is developing a new analysis paradigm—persistent and dynamic knowledge synthesis—that will provide continuous, automated synthesis of new knowledge and dynamic control of measurement systems contemporaneously with observed phenomena. Working on streaming data, this new capability will automate the current time-intensive manual analysis and interpretation steps and facilitate collaboration with scientists to optimize insight creation, decision making, analysis, and data capture to meet the needs of the discovery process in a timely manner. Example applications of these techniques could be *in situ* analysis of climate simulations or directing of data acquisition (e.g., ARM aircraft) based on real-time analysis and interpretation of instrument data.

Workflow Performance Modeling, Prediction, and Optimization

Workflows are taking an increasingly important role in orchestrating complex scientific processes in extreme scale and highly heterogeneous environments. However, current workflow performance cannot be reliably predicted, understood, and optimized. Sources of performance variability and, in particular, the interdependencies of workflow design, execution environment, and system architecture are not well understood. While there is a rich portfolio of tools for performance analysis, modeling, and prediction for single applications in homogenous computing environments, these tools are not applicable to workflows because of the number and heterogeneity of the involved workflow and system components and their strong interdependencies. PNNL is developing the capabilities to trace, validate, model, and predict workflow performance to inform better workflow design and enable runtime optimization (http://hpc.pnl.gov/IPPD/) by building on its leading research in systems architectures and performance modeling. The BER ACME project is one of the demonstrator use cases for this work.

Provenance and Reproducibility

Traditionally, provenance is used to explain the parentage of scientific results; however, as scientific workflows (both manual and computational) are increasing in complexity, provenance also can play a key role in supporting the tracing, validation, and reproduction of those workflows and their results to assess their performance, viability, and accuracy. In particular, the collection of provenance in extreme-scale environments and the reproducibility of scientific workflows present new research challenges. PNNL is continuing to develop the ProvEn provenance environment to address these challenges and derive actionable insights from the collected provenance records. This PNNL research focuses on the development of comprehensive provenance capture in extreme-scale environments, real-time provenance analysis, and provenance utilization to create enactable reproducibility records. The work is supported by a range of BER-, ASCR-, and PNNL-funded projects, including Climate Science for a Sustainable Energy Future (CSSEF, in the past), ACME, Integrated End-to-End Performance Prediction and Diagnosis for Extreme Scientific Workflows (IPPD), and AIM.

Collaborative Simulation and Analysis Environment

Scientific collaborations change regularly in their member composition, resource access, availability, and utilization. Keeping track of, sharing, and utilizing application, data, and resources in such distributed but collaborative settings are major challenges. The PNNL-developed Velo tool provides collaborative access mechanisms to these distributed resources, enabling users to register and utilize domain-specific data and metadata schema; register, share, and use data, existing tools, simulation codes, scripts, and workflows—as well as register, access, and use storage and computing and network resources. As such, Velo enables research collaborations to manage, jointly use their pooled resources, and share their research outcomes. Furthermore, this tool has a sophisticated security system that allows users to protect their research artifacts (data and tools) where required, but the system equally enables them to share artifacts at the appropriate time with the wider community (e.g., through publication). Velo also is linked to the ProvEn provenance system and thus can collect the required provenance and reproducibility information.

Acronyms and Abbreviations

3D	three dimensional
ACME	DOE Accelerated Climate Modeling for Energy
ADI	ARM Data Integrator
AdiFOR	automatic differentiation of Fortran software
AIM	PNNL Analysis in Motion initiative
ALCF	Argonne Leadership Computing Facility
ALM	ACME Land Model
AMR	adaptive mesh refinement
ANL	Argonne National Laboratory
API	application programming interface
ARM	CESD Atmospheric Radiation Measurement program
ASC	NNSA Advanced Simulation and Computing program
ASCEM	Advanced Simulation Capability for Environmental Management
ASCR	DOE Office of Advanced Scientific Computing Research
ATS	Advanced Terrestrial Simulator
BER	DOE Office of Biological and Environmental Research
BERAC	Biological and Environmental Research Advisory Committee
CADES	ORNL Compute and Data Environment for Science
CARR	OLCF Center for Accelerated Application Readiness
CESD	BER Climate and Environmental Sciences Division
CESM	Community Earth System Model
CLM	Community Land Model
COSIM	Climate, Ocean, and Sea Ice Modeling
CSSEF	CESD Climate Science for a Sustainable Energy Future
DAE	differential algebraic equation
DART	NCAR Data Assimilation Research Testbed
DOE	U.S. Department of Energy
DOE-RL	DOE Richland Operations
DOE-ORP	DOE Office of River Protection
DOI	digital object identifier
EDL	electrical double layer
EMSL	DOE Environmental Molecular Sciences Laboratory
EOS	equation of state
ESGF	NCAR Earth System Grid Federation
ESM	Earth system model
ESS	CESD Environmental System Science
FTIR	Fourier transform infrared spectroscopy
GPU	graphics processing unit
HEIS	Hanford Environmental Systems database
HPC	high-performance computing
IDEAS	DOE Interoperable Design of Extreme-scale Application Software
IDL	Interactive Data Language
IFRC	DOE Integrated Field Research Challenge
I/O	input/output
IP	intellectual property
IPPD	Integrated End-to-End Performance Prediction and Diagnosis for Extreme Scientific Workflows
LANL	Los Alamos National Laboratory
LBNL	Lawrence Berkeley National Laboratory
LES	large eddy simulation
LLNL	Lawrence Livermore National Laboratory
MADS	model analysis and decision support
MAPPER	Multiscale Applications on European e-Infrastructures
MATK	model analysis toolkit
MCMC	Markov Chain Monte Carlo
MCT	model coupling toolkit
MML	Multiscale Modeling Language
MODEX	model-driven experimentation and observation approach to predictive understanding
MOSART	MOdel for Scale Adaptive River Transport
MPAS	Model for Prediction Across Scales
NCAR	NSF National Center for Atmospheric Research
NERSC	DOE National Energy Research Scientific Computing Center
NGEE	TES Next-Generation Ecosystem Experiments
NNSA	DOE National Nuclear Security Administration
NSF	National Science Foundation
ODE	ordinary differential equation
OLCF	Oak Ridge Leadership Computing Facility
ORNL	Oak Ridge National Laboratory
PAWS	Process-based Adaptive Watershed Simulator
PDE	partial differential equation
PE	parameter estimation
PETSc	Portable, Extensible Toolkit for Scientific Computation
PFLOTRAN	massively parallel reactive flow and transport model for describing surface and subsurface processes
PHOENIX	PNNL Hanford Online Environmental Information exchange
PI	principal investigator
PMC	physics model coupler
PNNL	Pacific Northwest National Laboratory
PyART	Python ARM Radar Toolkit
RGCM	CESD Regional and Global Climate Modeling program
ROM	reduced-order model
SAChES	Scalable Adaptive Chain-Ensemble Sampling
SBR	CESD Subsurface Biogeochemical Research program
SC	DOE Office of Science
SciDAC	DOE Scientific Discovery through Advanced Computing
SDE	stochastic differential equation
SFA	DOE Scientific Focus Area (national laboratory research projects)
SNIA	sequential noniteritive algorithm
SUMO	SUrvival MOrtality experiments
TAO	Toolkit for Advanced Optimization
TES	CESD Terrestrial Ecosystem Science program
UQ	uncertainty quantification
WRF	Weather Research and Forecasting Model
xSDK	Extreme-Scale Software Development Kit