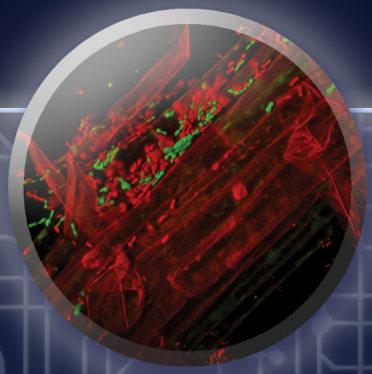
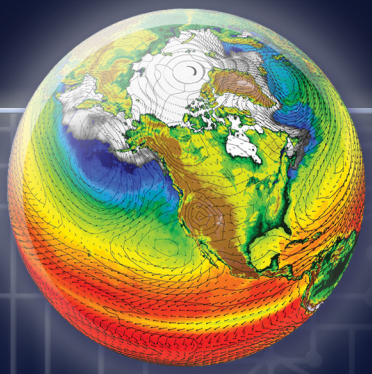
DOE/SC-0156





BER Virtual Laboratory: Innovative Framework for Biological and Environmental Grand Challenges

A Report from the Biological and Environmental Research Advisory Committee

Suggested citation for this report

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. science.energy.gov/ber/berac/reports/.

Cover

The Virtual Laboratory proposed for the Department of Energy's (DOE) Office of Biological and Environmental Research comprises three complementary components: integrated field laboratories; a Biosystems Frontier Network; and cyberinfrastructure, analytics, simulation, and knowledge discovery (CASK). Images on the cover represent how activities of each component (e.g., field measurements, plant and microbial characterizations, and modeling and simulation) would be part of an integrated framework for predictively understanding complex biological and environmental systems ranging from molecular to global scales.

Image credits: Microwave radiometer at a Maldives field site from the Atmospheric Radiation Measurement Climate Research Facility image gallery (flickr.com/photos/armgov). Fluorescent micrograph of rhizospheric bacteria colonized on root tissue from the Plant-Microbe Interfaces scientific focus area at Oak Ridge National Laboratory. Simulation from DOE's Parallel Climate Model project from the National Center for Atmospheric Research.

DOE/SC-0156

BER Virtual Laboratory: Innovative Framework for Biological and Environmental Grand Challenges

A Report from the Biological and Environmental Research Advisory Committee

U.S. Department of Energy

February 2013

Report available online at
science.energy.gov/ber/berac/reports/

Biological and Environmental Research Advisory Committee

Gary Stacey, Chair
University of Missouri

Dennis D. Baldocchi
University of California, Berkeley

Janet Braam
Rice University

Judith A. Curry
Georgia Institute of Technology

James R. Ehleringer
University of Utah

Susan Hubbard
Lawrence Berkeley National Laboratory

Andrzej Joachimiak
Argonne National Laboratory

L. Ruby Leung
Pacific Northwest National Laboratory

Gerald (Jay) Mace
University of Utah

Sabeeha Merchant
University of California, Los Angeles

Joyce E. Penner
University of Michigan

David A. Randall
Colorado State University

James T. Randerson
University of California, Irvine

Karin A. Remington
Arjuna Solutions

G. Philip Robertson
Michigan State University

William H. Schlesinger
Cary Institute of Ecosystem Studies

Martha A. Schlicher
Monsanto Company

Jacqueline V. Shanks
Iowa State University

Gaius (Gus) R. Shaver
Marine Biological Laboratory

Herman Shugart
University of Virginia

David A. Stahl
University of Washington

Judy D. Wall
University of Missouri

Warren M. Washington
National Center for Atmospheric Research

Minghua Zhang
State University of New York at Stony Brook

Huimin Zhao
University of Illinois, Urbana-Champaign

Designated Federal Officer

David Thomassen
U.S. Department of Energy Office of Biological and Environmental Research

BERAC Subcommittee Addressing Office of Science Charge

Janet Braam
Rice University

Judith A. Curry
Georgia Institute of Technology

Susan Hubbard*
Lawrence Berkeley National Laboratory

L. Ruby Leung
Pacific Northwest National Laboratory

Gerald (Jay) Mace
University of Utah

G. Philip Robertson
Michigan State University

Gary Sayler**
University of Tennessee

Gary Stacey
University of Missouri

Ray Wildung**
Pacific Northwest National Laboratory

Minghua Zhang
State University of New York at Stony Brook

---

*Special thanks to Susan Hubbard, who was instrumental in developing text and graphics for this report.
** Gary Sayler and Ray Wildung have rotated off BERAC since serving on this subcommittee.

Table of Contents

Executive Summary vii

1. Introduction 1

2. Integrated Field Laboratories 7

3. Biosystems Frontier Network 15

4. CASK—Cyberinfrastructure, Analytics, Simulation, and Knowledge Discovery 23

5. Strategic Demonstrations and Workshops 29

6. Checklist of Recommendations for Virtual Laboratory 31

Appendices 33

Appendix 1: Charge Letter 33

Appendix 2: Examples of Needed Technologies 35

Appendix 3: Bibliography 37

Acronyms ..... Inside back cover

Executive Summary

Identifying Capabilities Needed to Meet Grand Challenges

The Biological and Environmental Research Advisory Committee (BERAC) released in 2010 the *Grand Challenges for Biological and Environmental Research: A Long-Term Vision* report (science.energy.gov/~/media/ber/pdf/Ber_ltv_report.pdf). This document broadly addressed scientific opportunities and grand challenges in biological systems, climate, energy sustainability, computing, and education and workforce training. Several common science challenges were identified, including understanding complex systems science across scales, leveraging and expanding multidisciplinary research, advancing computing and mathematics capabilities, and assessing human impacts on the Earth system. The report also described the current fragmentation of science, technologies, and predictive capabilities among disciplines and the focus on studying mostly individual, scale-based system components. Such fragmentation leads to fundamental uncertainties about how coupled subsystems interact with each other and respond to environmental changes across different space and time scales. The lack of sufficient science-based capabilities to predict these interactions and responses hinders the ability to sustainably manage and mitigate energy and environmental problems. Associated with these science challenges, a number of knowledge and technology gaps were identified. In response to a follow-on request from the Department of Energy's (DOE) Office of Science, a BERAC subcommittee was formed to expand on the technology and tools most needed to support the biological and environmental research necessary for fulfilling the potential described in the Long-Term Vision report.

Developing a Virtual Laboratory to Understand Whole Systems

A cross-cutting innovation challenge emerged from the subcommittee's evaluation: the need to develop capabilities that enable a predictive understanding of complex, multiscale, coupled, and biologically based environmental systems behavior. Knowledge-based predictive models that describe a continuum of interactions occurring across the molecular (within or near a microbial or plant cell) to the regional and global scales are nonexistent. Achieving a predictive capability of this magnitude would be transformational, enabling a new class of solutions for environmental and energy challenges. After considering many options for technologies that would improve predictive understanding and capabilities, the subcommittee determined that the innovation most needed is a framework that allows seamless integration of multiscale observations, experiments, theory, and process understanding into predictive models for knowledge discovery. The envisioned strategy integrates existing assets within DOE's Office of Biological and Environmental Research (BER) and considers how they can best be expanded and exploited to achieve the opportunities articulated in the Long-Term Vision document.

BERAC thus recommends creating a BER **Virtual Laboratory** that would be developed by integrating and strategically expanding BER resources. A key goal of the Virtual Laboratory is to transition BER's research program from one associated with distributed datasets, specific process knowledge, and individual component models to one that provides a predictive understanding of key couplings and feedbacks among natural-system and

anthropogenic processes across scales. In other words, this transition involves moving beyond the investigation of "parts" to an understanding of integrated environmental systems behavior. As a virtual resource, the proposed laboratory would serve as a research integration tool to implement this framework via three key components: integrated field laboratories (IFLs); a Biosystems Frontier Network; and cyberinfrastructure, analytics, simulation, and knowledge discovery (CASK). Although these three components are described separately below and in more detail throughout this report, their integration is essential to the success of the Virtual Laboratory.

Integrated Field Laboratories

As key components in the Virtual Laboratory, IFLs should be integrated and expanded vertically (from the bedrock to the atmosphere) and geographically (across key geographic regions). The laboratories would capitalize on both existing BER field observatory investments—such as sites associated with the Atmospheric Radiation Measurement Climate Research Facility, AmeriFlux Network, subsurface biogeochemical field study sites, and Next-Generation Ecosystem Experiment program—and other observatories. These highly instrumented IFLs would traverse representative ecosystems and focus on understanding and scaling fundamental biogeochemical, microbial, and plant processes that drive planetary energy, water, and biogeochemical cycles. Rich in hypothesis-driven, process-level experimentation, the facilities would measure elemental, energy, and water transfer across mineral, biological, and atmospheric interfaces, including all ecological and climate processes that play a role in geochemical cycling. The IFL vision entails several vertically integrated sites in locations particularly significant to energy and climate and additional secondary sites where important biogeochemical, atmospheric, and biological processes can be quantified. An Instrument Incubator program tied to IFLs also is recommended to ensure development and implementation of needed measurement tools. This program would emphasize development of advanced field instrumentation. Data acquired at IFLs would be used in conjunction with the capabilities developed through the Biosystems Frontier Network and CASK to deepen scientific understanding of the major drivers and consequences of environmental change.

Biosystems Frontier Network

This network would integrate and expand technologies for microbial and plant physiology and phenomics to gain a better understanding of organism and community phenotypic expressions within complex and highly dynamic natural environments. The development of *in situ* and nondestructive "omics" approaches is needed to identify biosphere stress-specific signatures that reflect the impact of climate and environmental changes under natural conditions and in near-real time. New quantitative measurement capabilities will enable nondestructive analyses across expansive temporal (nanosecond to years) and spatial (subnanometer to kilometer) scales and in diverse environments. Also needed are bioimaging innovations that (1) allow multifunctional probes and labeling chemistries to provide contrasts across multiple imaging platforms; (2) advance image analysis, registration, and feature measurement algorithms to quantify images; and (3) integrate and relate information arising from different methods. These new *in situ* measurement and bioimaging capabilities, which include improved techniques for examining cellular molecules, metals, and ions, must be comprehensive and cost-effective to enable study of the full diversity of organisms present in different natural environments. The subcommittee recommends development of a distributed research and instrumentation resource built upon the core expertise of existing BER user facilities such as the DOE Joint Genome Institute and Environmental Molecular Sciences Laboratory. Such a resource

will enable researchers to quantify function (e.g., protein activities, metabolic fluxes, physiology, and biogeochemistry) in the context of complex, diverse, and highly dynamic environments.

CASK—Cyberinfrastructure, Analytics, Simulation, and Knowledge Discovery

CASK would provide the computational infrastructure needed to integrate disparate and multiscale measurements, theory, and process understanding into predictive models, ultimately creating new knowledge that can be used to develop energy and environmental solutions. CASK data assimilation and simulation tools would aid development of the next generation of advanced theory and the integration of information spanning molecular to global scales, enabling predictions about environmental changes that will inform policy. Building upon and integrating BER's existing knowledge discovery infrastructure (such as the DOE Systems Biology Knowledgebase and Carbon Dioxide Information Analysis Center), CASK would develop strategies for linking heterogeneous databases and for federation and exchange of information obtained from IFLs, the Biosystems Frontier Network, and other resources. The resulting environment for multiscale knowledge discovery would provide innovative capabilities in distributed data discovery, visualization, analysis, and uncertainty quantification. Also envisioned for CASK is the development of advanced system component models. Examples of needed improvements include incorporating cell function into reactive transport models, coupling watershed biogeochemical simulators to land models, and advancing DOE climate models.

Outlining a Path for Virtual Laboratory Development

This report lays out a framework for technological innovation within the context of the BER Virtual Laboratory. It also identifies the areas of greatest need and opportunity for transitioning BER science from observational, single-system studies to research focused on predictive understanding of the key components governing complex biological and environmental systems. To fully implement the action items described herein, BERAC provides several component-specific recommendations within this report. The committee also proposes the following workshops and strategic demonstrations be initiated in the near term to orchestrate BER Virtual Laboratory development.

Near-Term Workshops

- **Environmental Observatories Workshop** to identify and prioritize opportunities for expanding and integrating BER and other community observatory investments, both vertically (from the bedrock to the atmosphere) and geographically (across key geographic regions).
- **Biosystems Frontier Network Workshop** to explore the development of a distributed research and instrumentation resource built upon the core expertise of existing BER user facilities. Such a resource would enable quantitative observations and *in situ* measurements of biomolecules that can be linked to function in the context of complex and diverse physical, chemical, and biological environments.
- **Multiscale Simulation and Data Assimilation Workshop** that calls on diverse experts from the scientific community to create a detailed, prioritized roadmap for developing multitype, multiscale data assimilation, simulation, and knowledge discovery capabilities needed to address cross-cutting BER grand challenges. This workshop should consider the technical, mathematical, theoretical, and computational challenges associated with data assimilation and strategies for fully leveraging BER's myriad databases and knowledgebases to gain a predictive understanding of multiscale phenomena.

- **Implementation Workshop** to synthesize and prioritize recommendations resulting from the three previous, component-specific workshops and to develop a roadmap for creating the BER Virtual Laboratory.

Strategic Demonstrations: First Steps Toward the Virtual Laboratory

After the workshop series, BERAC recommends initiating strategic demonstrations to advance and test one to three prototype designs for the Virtual Laboratory. These strategic demonstrations are envisioned to guide the development and integration of the full-scale Virtual Laboratory by formulating and testing a prototype in conjunction with selected field laboratories, Biosystems Frontier Network capabilities, and CASK components. The strategic demonstrations also aim to identify the benefits and limitations of the Virtual Laboratory concept. Guided by hypothesis-driven scientific questions that require predictive understanding of multiscale environmental phenomena, these demonstrations will explicitly define how the Virtual Laboratory will leverage and integrate BER and other community resources.

1 Introduction

Building on BERAC Long-Term Vision Opportunities

This document—developed in response to a charge delivered to the Biological and Environmental Research Advisory Committee (BERAC)—expands on the technology needs identified in a 2010 BERAC report titled *Grand Challenges for Biological and Environmental Research: A Long-Term Vision*. The 2010 report, hereafter referred to as the Long-Term Vision document, described research "… that can put society on a path to achieve the scientific evidence and predictive understanding needed to inform decision making and planning to address future energy needs, climate change, water availability, and land use." The charge, given by Dr. William Brinkman, director of the Department of Energy's (DOE) Office of Science, is to:

- Expand on the development and use of new tools mentioned in the Long-Term Vision report.
- Identify the development and use of new tools and their linkage to existing or new user facilities.
- Identify linkages between new and existing resources and to diverse scales of time and space.
- Expand on the concepts of virtual laboratories and collaborative tools, including a discussion of how to facilitate these concepts and interactions.

The BERAC Long-Term Vision report broadly addressed scientific opportunities and grand challenges in biological systems, climate, energy sustainability, computing, and education and workforce training. Several common science challenges were identified, including understanding complex systems science across scales, leveraging and expanding multidisciplinary research, advancing computing and mathematics capabilities, and assessing human impacts on the Earth system. Many associated technology gaps were identified, a subset of which is provided in Appendix 2 (see p. 35). In response to Dr. Brinkman's charge, BERAC formed a subcommittee that held a series of teleconferences, gathered input from disciplinary-based focus groups and other BERAC members, and discussed the developing ideas with the full BERAC committee during meetings.

Enabling Multiscale Knowledge Discovery: The Virtual Laboratory

Through these deliberations, a cross-cutting innovation challenge emerged: the need to develop capabilities that enable a predictive understanding of complex, multiscale, coupled, and biologically based environmental systems behavior. Knowledge-based predictive models that describe a continuum of events occurring from the molecular (within or near a microbial or plant cell) to the regional and global scales are nonexistent. Achieving a predictive capability of this magnitude would be transformational, enabling a new class of solutions for environmental and energy challenges (see Fig. 1.1. Investigating Biological and Environmental Interactions Across Scales, p. 2). Such a transformation would entail unprecedented integration of *in situ* hypothesis-driven experimentation and observations; technological advances across multiple scales of space, time, and system organization; and major advances in theory, mathematics, and computation associated with multiscale models. Although necessary for the informed development of scientifically sound policy in the coming decades, the present state of science, technologies, and predictive capabilities is fragmented across disciplines and

individual system components that operate over a limited scale range. Such fragmentation leads to fundamental uncertainties about how coupled subsystems interact across space and time scales. For instance, microbial and plant community activity and hydrogeochemical parameters, which are both predictors and indicators of biology (as well as forcing agents), interact with each other and the environment to govern critical Earth system processes. They respond to atmospheric inputs of energy and water to control the production of food and biofuel feedstocks and regulate the fluxes of clean water in the subsurface and greenhouse gases to and from the atmosphere. The lack of sufficient science-based capabilities to predict these community and environmental interactions and their response to change over relevant space and time scales hinders the ability to sustainably manage and mitigate energy and environmental challenges.

After considering many options for technologies that would improve predictive understanding and capabilities, the subcommittee determined that the innovation most needed is a framework enabling seamless integration of multiscale observations, experiments, theory, and process understanding into predictive models for knowledge discovery. Such a framework would require consideration of coupled biological, hydrological, geochemical, and physical processes across a wide range of spatial and temporal scales important to the coupled evolution of these processes as the Earth changes in the coming decades. The

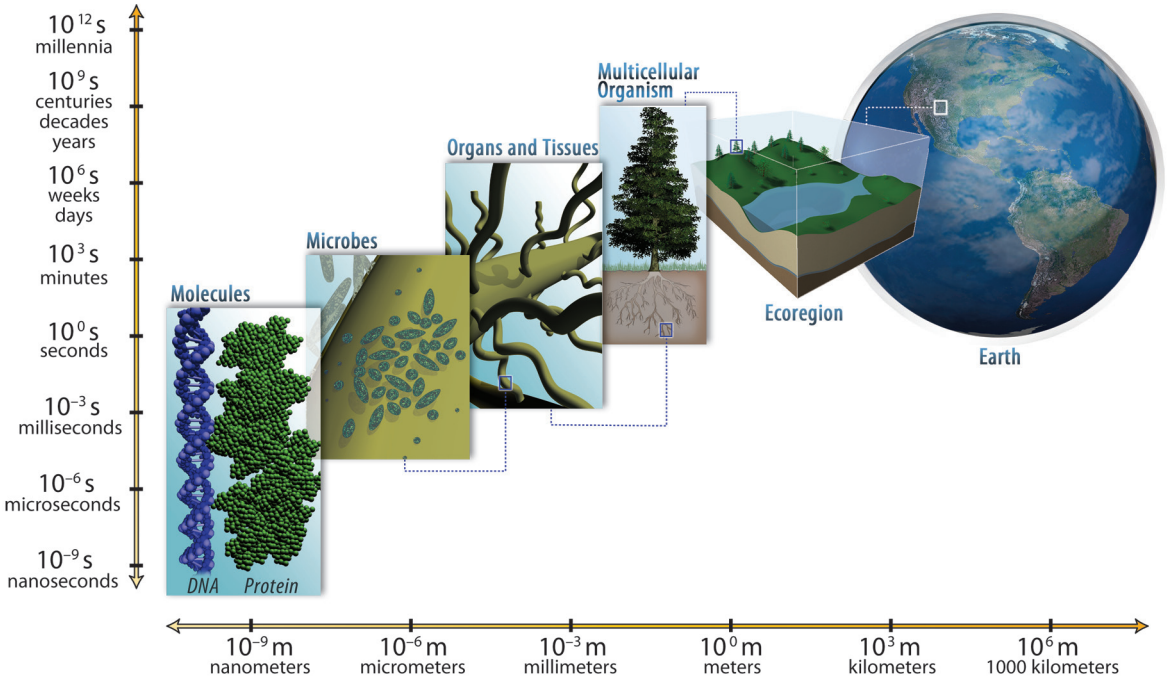


Fig. 1.1. Investigating Biological and Environmental Interactions Across Scales. Understanding how complex biological and environmental systems will respond to and affect critical Earth system processes requires measuring, simulating, and integrating biological, chemical, and physical components and their interactions across vast spatial and temporal scales—from subnanometers to kilometers and nanoseconds to millennia. Shown here is the example of moving from an initial view of molecules inside a cell, to cellular components, whole cells, the cells of a tissue interacting with adjacent cells, the tissue as part of an organism (e.g., a tree), the organism within a geographical region of a continent, and finally to a view of the whole Earth. As computational models become refined and increasingly precise at each scale, the question of how a model can interact meaningfully with those at adjacent scales will present opportunities to gain a deeper system understanding. [Image from p. 46 of the *Grand Challenges for Biological and Environmental Research: A Long-Term Vision* report.]

envisioned strategy incorporates existing assets within DOE's Office of Biological and Environmental Research (BER) and considers how they can best be advanced and exploited to realize the opportunities articulated in the Long-Term Vision document.

BERAC thus recommends the development of a BER **Virtual Laboratory**. As a virtual resource, the proposed laboratory would serve as a research integration tool to implement the envisioned research framework via three key complementary components: integrated field laboratories (IFLs); a Biosystems Frontier Network; and cyberinfrastructure, analytics, simulation, and knowledge discovery, or CASK (see Fig. 1.2. Biological and Environmental Research Virtual Laboratory, this page).

Integrated Field Laboratories. By integrating and expanding existing BER field observatories, IFLs would advance efforts to understand key biogeochemical, biological, and hydrological processes that influence ecosystem behavior from the bedrock to the atmosphere under varied environmental conditions and landscapes. These highly instrumented laboratories also would serve as validation points for deepening scientific understanding of the major drivers and consequences of environmental change arising from both natural variability and human activities.

Biosystems Frontier Network. This network would coordinate with BER user facilities, DOE national laboratories, and other collaborators to improve understanding of the dynamic functions and physiological responses of plants and microbes across scales and in complex, diverse environments (see Sidebar 1, BER User Facilities Provide Core Capabilities for the Virtual Laboratory, p. 5). Achieving this understanding will require (1) advanced tools for measuring and observing biomolecules, metals, and ions; (2) *in situ* and nondestructive "omics" approaches to

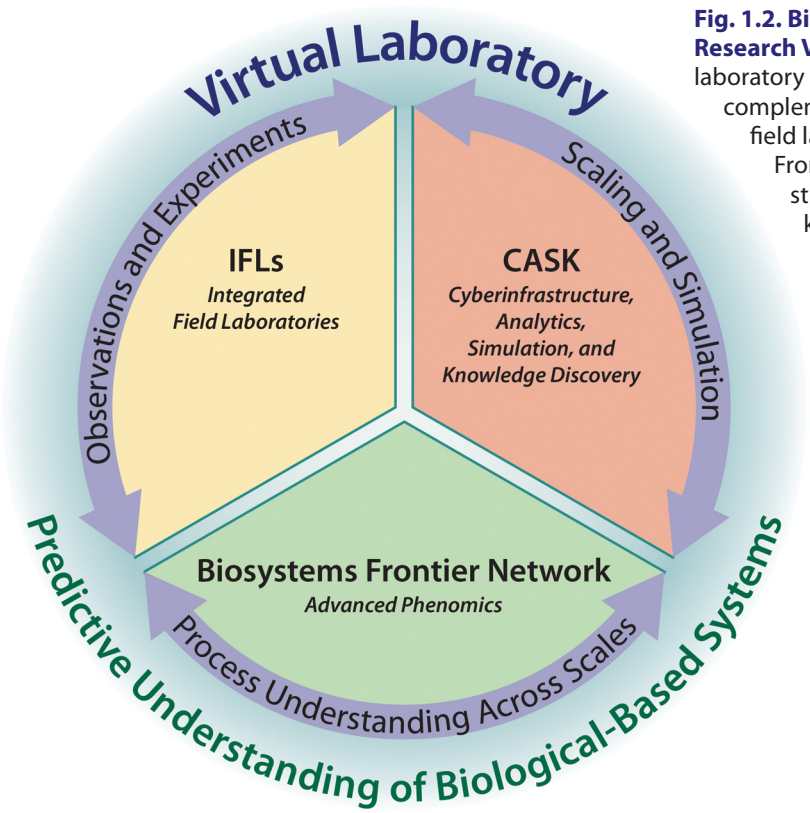


Fig. 1.2. Biological and Environmental Research Virtual Laboratory. The proposed laboratory consists of three integrated and complementary components: integrated field laboratories (IFLs); the Biosystems Frontier Network; and cyberinfrastructure, analytics, simulation, and knowledge discovery (CASK).

identify biosphere stress-specific signatures that reflect environmental changes in near-real time; and (3) improved bioimaging tools.

CASK—Cyberinfrastructure, Analytics, Simulation, and Knowledge Discovery. This computational infrastructure would integrate disparate and multiscale measurements, theory, and process understanding into predictive models, ultimately creating new knowledge for energy and environmental solutions. Building on BER's existing resources for knowledge discovery (e.g., the DOE Systems Biology Knowledgebase and Carbon Dioxide Information Analysis Center), CASK would link heterogeneous databases and develop advanced data assimilation and simulation tools for hypothesis testing and integration of multiscale, multitype data (including information from IFLs and the Biosystems Frontier Network).

Moving Toward Predictive Understanding of Complex Systems

By expanding and integrating existing BER and other community science resources, the Virtual Laboratory will facilitate the transition of BER's research program from one associated with distributed datasets, specific process knowledge, and individual models to one that provides a predictive understanding of key couplings and feedbacks among natural-system and anthropogenic processes across scales. In other words, this transition involves moving beyond the investigation of "parts" to an understanding of integrated environmental systems behavior. Although each of the three Virtual Laboratory components is important, it is their integration that will enable the transformative predictions of complex biological system functioning that can then guide management of and solutions to energy and environmental problems.

Establishing and maintaining the Virtual Laboratory represent both a challenge and new collaboration opportunities across the BER science portfolio. Accessible remotely, the Virtual Laboratory will provide important enabling developments and tools and will serve as a gateway to advanced resources and databases. Excellent examples of this strategy are the DOE-funded synchrotrons and their "virtual x-ray facilities," which were pioneered several years ago and have been applied with great success in the structural biology field. This highly feasible approach has transformed how users interact with the instrumentation and resources, with more than 90% of users at several synchrotrons accessing the facility remotely. This model can be further expanded and implemented in other areas of biological and environmental research.

The three components of the BER Virtual Laboratory are described in detail in Chapters 2–4, which also include recommendations associated with each. The component descriptions are followed by a discussion of actions needed to facilitate integration of the three into an advanced Virtual Laboratory. These actions include initiating strategic demonstrations to test prototype designs for the laboratory and organizing several community workshops to facilitate component development and implementation. The final chapter summarizes overarching recommendations for the demonstrations, workshops, and laboratory components.

Sidebar 1

BER User Facilities Provide Core Capabilities for the Virtual Laboratory

The science supported by the Department of Energy's (DOE) Office of Biological and Environmental Research (BER) is rapidly changing, presenting both opportunities and challenges associated with the goal of understanding complex biological and environmental systems across many spatial and temporal scales. To address these challenges and keep pace with the science, instruments must be continually improved and upgraded and new tools developed. These efforts occur within a distributed system of BER-supported laboratories but are particularly focused within BER's scientific user facilities. These facilities thus can provide the core expertise and technology needed to develop the BER Virtual Laboratory envisioned in this report.

BER's Climate and Environmental Sciences Division supports two important user facilities: the Atmospheric Radiation Measurement (ARM) Climate Research Facility and the Environmental Molecular Sciences Laboratory (EMSL). The ARM Climate Research Facility (**www.arm.gov**) seeks to improve scientific understanding of the fundamental physics related to the interactions between clouds and radiative feedback processes in the atmosphere. Its infrastructure includes a data archive, highly instrumented ground stations, mobile resources, and aerial vehicles to continuously measure cloud and aerosol properties and provide data products that promote the advancement of climate models. EMSL (**www.emsl.pnnl.gov**) offers experimental resources for discovery and technological innovation in the environmental molecular sciences. It provides cutting-edge technologies in mass spectrometry, nuclear magnetic resonance spectroscopy, imaging, and bioreactor capabilities, as well as computational resources to support these tools. The facility also has important molecular-scale capabilities to support atmospheric research and is envisioned to play a significant role in the Biosystems Frontier Network (see Chapter 3, p. 15) by addressing critical questions concerning biological and environmental complexity.

The Biological Systems Science Division supports the DOE Joint Genome Institute (JGI) and a structural biology infrastructure consisting of DOE synchrotron and neutron beam facilities for biological investigations at the atomic or molecular level. This infrastructure (**genomicscience.energy.gov/userfacilities/structuralbio.shtml**) represents a distributed resource of nine experimental stations at six laboratories that each offer unique expertise and instrumentation for examining questions about the structure of matter. The organization of this infrastructure is perhaps a model for how the Biosystems Frontier Network might operate on a larger scale. However, addressing the challenges of 21st century biology (see Sidebar 4, p. 18) will require a level of coordination and integration that exceeds current practice.

DOE JGI (**www.jgi.doe.gov**) is the only federally funded high-throughput genome sequencing and analysis facility dedicated to genomes of nonmedical microbes, microbial communities, plants, fungi, and other targets relevant to DOE missions in energy, climate, and the environment. The facility offers users access to massive-scale DNA sequencing to underpin modern systems biology research and provides fundamental data on key genes that may link to biological functions. A recent workshop outlined an exciting and innovative vision for DOE JGI's future (**genomicscience.energy.gov/userfacilities/jgi/futuredirections/**) that is wholly consistent with the goals of the BER Virtual Laboratory, especially when integrated fully with the other initiatives described in this report.

2 Integrated Field Laboratories

Leveraging Existing Field Observations and Experiments

The Biological and Environmental Research Advisory Committee's (BERAC) Long-Term Vision report emphasized multiscale complexity as a major challenge for the 21st century. Scientists face the demanding tasks of integrating observations and process understanding across vast temporal and spatial scales, maintaining and connecting very large datasets, and predicting feedbacks and emergent properties common to complex biophysical systems that respond to and influence environmental changes.

Field study sites have long served as natural laboratories for quantifying biological, hydrological, atmospheric, and geochemical processes important for understanding global change; sequestering contaminants; and securing fresh water, food, and energy crops. Subsurface, land surface, and atmospheric sites provide observations and experiments that test and develop hypotheses and models, ultimately creating new knowledge. The importance of field study sites is underscored by Office of Biological and Environmental Research (BER) investments in several distinct suites of environmental observatories distributed across and, in some cases, beyond the continental United States, including:

- Integrated field research challenge (IFRC) study sites that focus on subsurface biogeochemical processes and system functioning from the land surface to the bedrock.
- Next-Generation Ecosystem Experiment (NGEE) studies, including one under development in the Arctic and one envisioned for the Tropics. These projects are designed to quantify feedbacks between terrestrial ecosystems and climate.
- AmeriFlux sites that measure the exchange of carbon dioxide (CO_2) and other trace gases between the land surface and atmosphere.
- Sites operated by the Atmospheric Radiation Measurement (ARM) Climate Research Facility that measure aerosol, cloud, and precipitation properties and atmospheric dynamics important for developing and validating high-resolution climate models.

These sites, as well as those proposed in this report, can serve as testing grounds for developing sensors that collect data for model parameters and state variables at the scales needed to support modeling efforts. Although BER supports a number of measurement activities, they are not integrated at a true systems level and do not readily couple real-time data feeds to models. The potential exists to more closely couple field measurements to models that can then be used to better define the questions to be addressed in a field laboratory.

BER investments in field measurements are complemented by controlled experiments and additional observations conducted both within and outside BER. These activities and resources include controlled laboratory studies of cloud and aerosol formation processes at the Environmental Molecular Sciences Laboratory (EMSL), field observatories supported by the National Science Foundation (NSF), and global satellite measurements collected by the National Aeronautics and Space Administration (NASA) and National Oceanic and Atmospheric Administration. The controlled experiments at EMSL, for example, have changed fundamental understanding of the phase of secondary organic aerosols, knowledge that has since been confirmed by ARM field measurements. NSF's National Ecological Observatory Network (NEON) is designed

to detect continental-scale environmental change via a geographically dispersed system of 20 sensor towers in natural habitats. The agency's Long-Term Ecological Research (LTER) program uses observations and experiments to formulate and test hypotheses about long-term ecological change in 26 distributed sites, and NSF's Critical Zone Observatories explore geological and Earth surface processes at four locations. BER and NSF field observatories have been invaluable for understanding important processes that occur at specific sites and scales. In addition, the ARM Climate Research Facility and NASA's Earth Science program have established a natural partnership, with ARM sites providing critical validation information for NASA satellites whose data, in turn, provide regional context for ARM ground-site measurements.

Challenges of Linking Multiscale Ecosystem Processes to Climate

Several significant gaps limit the ability to use current field-based measurements to meet the grand challenges of predicting environmental change and understanding how it influences biological system functioning across a vast range of relevant spatial scales. Flows of carbon, nitrogen, water, and other biogeochemical elements depend fundamentally on interactions occurring at multiple scales across different system components, from the bedrock to the atmosphere. Yet too little is known, for example, about how plant and microbial community structures affect ecosystem function and climate (see Sidebar 2, Cells in Ecosystems, p. 9). Of particular and new importance is the need for a high-resolution understanding of ecosystem processes and outcomes. Climate models of the coming decades will resolve scales down to several kilometers, allowing prediction of small-scale behaviors. However, these models will be especially sensitive to fine-scale atmospheric turbulence. Atmospheric measurements, therefore, must be of sufficient temporal and spatial resolution to inform high-resolution models that have data assimilation systems uniquely developed to ingest these measurements. Results from the coupling between measurements and models could then be used to develop the predictive understanding that can be incorporated into global climate models. As described in Sidebar 3, Aerosol and Cloud Interactions, p. 10, integrated field laboratories (IFLs) can be developed to provide the needed information about local-scale subsurface-surface coupling and surface attributes and processes that affect regional cloud formation, including aerosol formation.

Although ongoing research focuses on exploring how interactions within individual ecosystems lead to particular environmental outcomes, tackling how these interactions and outcomes scale up from cells to watersheds, airsheds, and regions amid a rapidly changing climate is a huge challenge. A key reason for this challenge is the limited variety of measurement suites associated with existing geographically distributed observatories. For example, AmeriFlux sites measure CO_2 fluxes from soil respiration but not the microbial and rhizospheric interactions that drive respiration, nor specific contributions by the plant community. IFRCs, on the other hand, measure subsurface processes at locations different from the AmeriFlux sites without linking them to soil and aboveground processes. Since neither of these networks provides the information needed to link ecosystem processes to regional climate, interactions and feedbacks occurring from the bedrock through the canopy cannot be fully articulated. Moreover, understanding human-driven perturbations resulting in large-scale environmental changes also depends on both model development and the availability of appropriate observations. IFLs could provide important validation points for improving this understanding.

Another challenge is the difficulty of quantifying the functioning of molecules, microbes, and microbial and plant communities *in situ* and under varied environmental conditions, especially those associated with climate change. Also critical is the integration of observatory data across scales and sites to provide a predictive geographic understanding enabled by

Sidebar 2

Cells in Ecosystems

Biogeochemical flows in terrestrial ecosystem models historically have been represented as the outcome of environmental influences on "black boxes" that represent large taxonomic groups—most commonly plants and microbes. More recently, these boxes have been usefully disaggregated into functional groups. For plants, the groups include legumes, grasses, and trees. For microbes, they consist of autotrophs (e.g., nitrifiers) and heterotrophs with and without specialized metabolisms (e.g., nitrate or sulfate respirers). The power of including greater functional and even taxonomic resolution in models is readily apparent from models of carbon flow in plant communities. For example, knowledge of individual species and their phenologies within complex plant communities allows predictions of drought effects on primary production, storm impacts on hydrologic nutrient loss, and many other important ecosystem fluxes affected by slow or episodic environmental changes to which plants respond.

The same is not true for microbes except in the broadest functional sense. This is largely because scientists until only recently have had a very poor understanding of the effective complexity of soil microbial communities—whether they inhabit the plant rhizosphere, surface of soil mineral particles, interior of soil aggregates, or deep subsurface environments such as groundwater. However, the understanding of microbial communities is expanding rapidly. Advances in sequencing applied to whole communities, such as the DOE Joint Genome Institute's Great Prairie project (Jansson 2011) and the Rifle, Colo., subsurface biogeochemistry project (Wrighton et al. 2012), are providing metagenomic information on the relative importance of various taxonomic groups in different soil habitats. This information allows the inference of function that can be confirmed via experimentation and then eventually incorporated into better ecosystem models. These improved models will enable researchers to predict the fluxes influenced by microbes with the same precision available for similar plant studies.

Trace gas fluxes illustrate this point. Nitrous oxide (N_2O), for example, is an important greenhouse gas whose global flux is mostly of microbial origin. The microbial processes responsible for N_2O production in soil (nitrification and denitrification) are known, as are, in general, the environmental factors that most affect production (i.e., soil moisture, temperature, pH, redox status, and carbon and nitrogen availability). However, models of N_2O flux are notoriously poor at predicting changes in day-to-day fluxes in different ecosystems, likely related to taxonomic differences in the nitrifier and denitrifier communities responsible for N_2O production. New metagenomic information suggests that the taxonomic diversity of these groups differs by ecosystem and vertically by habitats within ecosystems. Different taxa likely have been selected by and respond to environmental influences differently. Understanding these differences and how they are related to the distribution of different nitrifiers and denitrifiers from the bedrock to the soil surface will be essential for accurately modeling ecosystem N_2O fluxes in response to environmental change.

References

Jansson, J. 2011. "Towards 'Tera-Terra': Terabase Sequencing of Terrestrial Metagenomes," *Microbe* **6**(7), 309–15.

Wrighton, K. C., et al. 2012. "Fermentation, Hydrogen, and Sulfur Metabolism in Multiple Uncultivated Bacterial Phyla," *Science* **337**(6102), 1661–65. DOI: 10.1126/science.1224041.

Sidebar 3

Aerosol and Cloud Interactions

Aerosol effects on clouds consist of three sequentially linked controls. First, increased numbers of aerosols in cloudy volumes provide additional locations for droplet nucleation and, all else being equal, result in clouds that have more but smaller droplets and thus are more reflective to solar radiation (Twomey 2007). Second, because smaller droplets do not collide and coalesce as efficiently, increased numbers of them are hypothesized to hinder precipitation (Albrecht 1989). Third, precipitation suppression then leads to longer-lived clouds that reflect back to space the solar radiation that usually would have been absorbed at the surface had the clouds been shorter lived. This straightforward progression of aerosol effects is easily understood and generally accepted. The problem, however, is that there is scant evidence that these effects actually have a measurable impact on overall system behavior.

The boundary-layer clouds that aerosol perturbations influence most strongly in the manner described above exist in a specific state as one possible way the atmosphere transports energy between the surface, boundary layer, and free troposphere (Stevens and Feingold 2009). However, a large, if not infinite, number of coupled pathways are available to the climate system to accomplish this energy transport. Perturbations to one characteristic of the system, such as an increase in sulfate aerosol in a region, could be buffered by changes to other aspects of the system so that the net effect on climate may be small. In any event, hypothesizing climate impacts via the simplistic one-dimensional arguments presented above is not sufficient for understanding and predicting system behaviors.

The means by which this issue can be best addressed illustrate clearly the challenges and opportunities described in this report. Although improved observations of aerosols, clouds, and precipitation are needed from integrated field laboratories, observations alone are insufficient. Advanced use of data, such as combining information from diverse data sources and utilizing data-model integration techniques, is needed to understand how aerosol loadings perturb the microphysical processes and associated impacts to fluxes of energy and condensed water within a cloud and cloud systems. Integrating advanced modeling (including cloud-resolving simulations) with observations and uncertainty quantification techniques can facilitate exploration of the responses of not only cloud systems but also the coupled climate system. Within the cyberinfrastructure, analytics, simulation, and knowledge discovery (CASK) component of the BER Virtual Laboratory, information can then be converted to knowledge by exploring large, diverse datasets and gaining insights from numerical models from which general conclusions can be drawn.

References

Albrecht, B. A. 1989. "Aerosols, Cloud Microphysics, and Fractional Cloudiness," *Science* **245**(4923), 1227–30.

Stevens, B., and G. Feingold. 2009. "Untangling Aerosol Effects on Clouds and Precipitation in a Buffered System," *Nature* **461**, 607–13.

Twomey, S. 2007. "Pollution and the Planetary Albedo," *Atmospheric Environment* **41**, 120–25.

simulation capabilities that are currently lacking. Ultimately, what is most needed is a tighter integration of field laboratory experiments and observations with new phenomic knowledge provided by the second component of the BER Virtual Laboratory, the Biosystems Frontier Network. This knowledge and information will be integrated and synthesized by the cyberinfrastructure, analytics, simulation, and knowledge discovery (CASK) component of the laboratory.

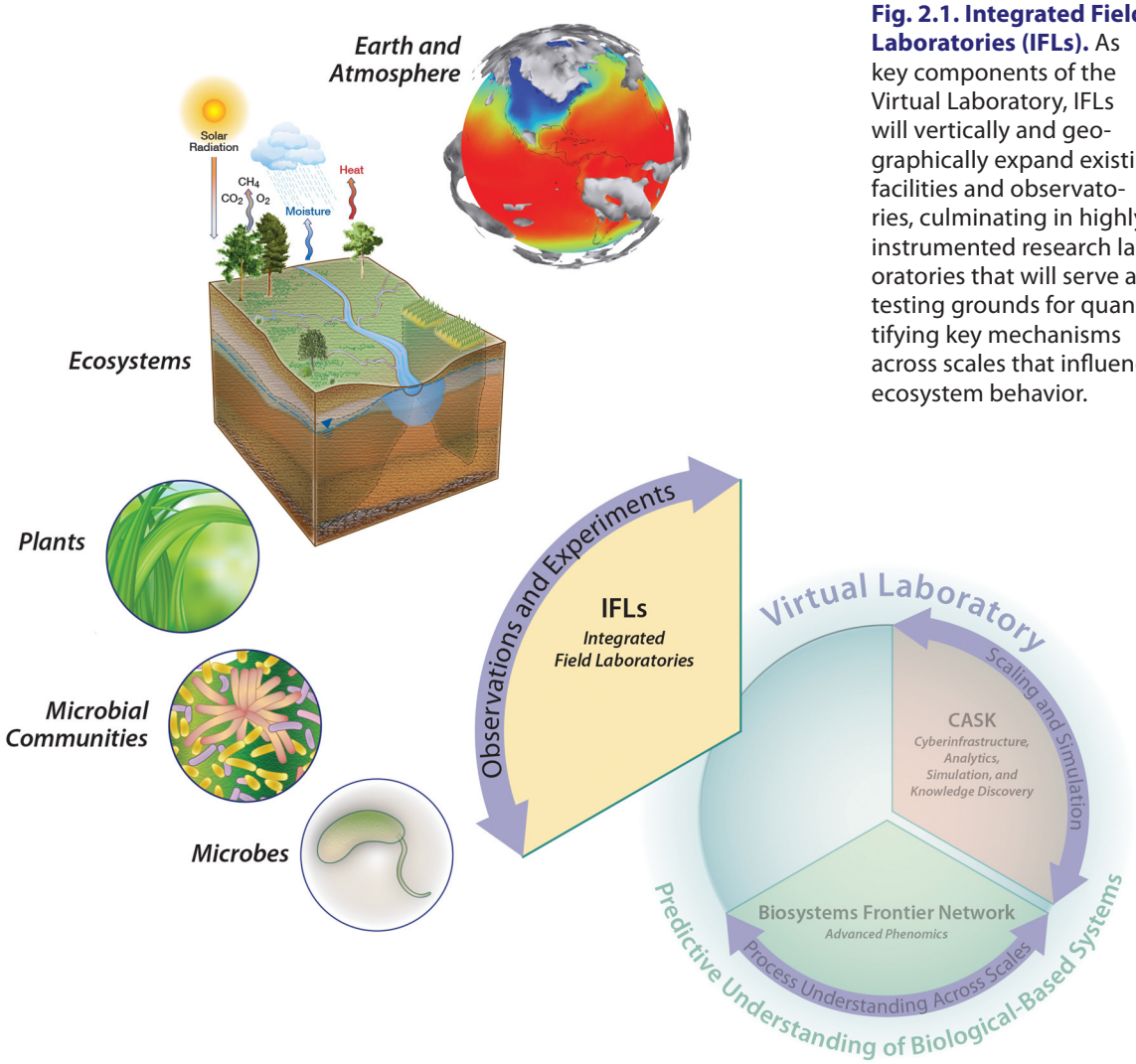


Fig. 2.1. Integrated Field Laboratories (IFLs). As key components of the Virtual Laboratory, IFLs will vertically and geographically expand existing facilities and observatories, culminating in highly instrumented research laboratories that will serve as testing grounds for quantifying key mechanisms across scales that influence ecosystem behavior.

CASK will provide a means for organizing and confronting data using newly developed models capable of simulating both vertically and geographically within and across field laboratories. The ultimate goal is to turn data and information into new knowledge and understanding.

Understanding and Scaling Key Processes via IFLs

As a key component of the Virtual Laboratory, BERAC proposes integrating and expanding IFLs, both vertically (from the bedrock to the atmosphere) and geographically (across key geographic regions). IFLs necessarily would include multiple, highly instrumented research field laboratories in representative ecosystems (either co-located or in close proximity) focused on understanding and scaling fundamental biogeochemical, plant, and microbial processes that drive planetary energy, water, and biochemical cycles (see Fig. 2.1. Integrated Field Laboratories, this page). Rich in hypothesis-driven, process-level experimentation, IFLs would measure elemental, energy, and water transfer across mineral, biological, and atmospheric interfaces, including all ecological and climate processes that play a role in geochemical cycling (e.g., the atmosphere, plants, soils and sediments, vadose zone, groundwater,

hyporheic zone, and surface waters). The resulting observational data would provide objective functions and a framework to quantify how key mechanisms and processes at various scales couple to control essential system behaviors.

Specific IFL Recommendations

Field laboratories with deep vertical integration should be established in a small number of landscapes particularly important to energy and climate. Research would center on hypothesis-driven questions focused especially at boundaries across scales (such as those described in Sidebars 2, p. 9, and 3, p. 10), ultimately providing the predictive understanding and capacity to model key processes involved in environmental change related to energy and climate systems. Building upon BER's NGEE concept, the vertically integrated IFLs would include deeper subsurface and atmospheric measurements. Vertically integrated research would be complemented and extended by controlled laboratory studies and research at a separate set of geographically dispersed field sites that also build upon BER observatory investments and modeling of the driving forces of change. These latter sites would provide, for specific processes, the opportunity to test models developed at the vertically integrated sites and extend the range of environments over which key processes are understood. This capability would provide a geographic resolution not possible with a limited number of vertically integrated sites. To implement this strategy, the following three steps are recommended:

Identify Opportunities for Leveraging BER Investments to Develop Vertically Integrated Laboratories. The identified IFLs would be located in environments representative of large, rapidly changing regions or areas strategically important for the bioeconomy. These laboratories would provide the capacity to scale across biophysical boundaries spanning from the deep subsurface to the upper atmosphere, involving both multiple media (e.g., water, minerals, soil, and air) and taxa (e.g., microbes, plants, and other organisms such as plant pathogens and symbionts). This research would advance understanding of the fundamental flows of energy, water, carbon, nitrogen, and other critical elements that interact to affect large-scale climate and energy feedbacks, environmental health, and ecosystem productivity and examine how human activities perturb natural cycles. The vertically integrated sites could be developed by extending the measurement suites for NGEE Arctic and other existing field sites and by establishing several new vertically integrated sites in conterminous U.S. regions that are important to climate and bioenergy and that take advantage of existing infrastructure (such as the Great Lakes, Midwest, Southern Plains, or western U.S. regions).

Strategically Identify Geographically Dispersed Sites with the Necessary Subsurface, Land Surface, and Atmospheric Components. These secondary sites would include BER investments in existing resources such as the ARM Climate Research Facility, subsurface biogeochemical sites, and AmeriFlux Network and could perhaps be expanded to leverage NSF investments in LTER, NEON, and Critical Zone Observatories. Specific linkages with existing BER research are described below.

- Subsurface biogeochemical IFRCs would provide the necessary experimental systems to quantify *in situ* biogeochemical processes occurring from the land surface to the bedrock, in surface and oceanic waters, and from cell to watershed scales. Such research also would enable investigation of how these processes are expected to change as a function of climate and land use. Leveraging and extending BER IFRCs traditionally focused on contaminant flow and transport, these centers will examine biogeochemical processes occurring in the context of hydrological systems to quantify biogeochemical watershed functioning that influences carbon cycling, climate change, and biofuel crop sustainability. Given their design to develop robust understandings of linked biological and hydrochemical processes, these sites provide unique opportunities to

characterize and model the hierarchical function of complex environmental systems spanning heterogeneous molecular processes to kilometer-scale phenomena. Subsurface biogeochemistry field sites would be strategically expanded to include important hydrological, biogeochemical, and ecosystem types or marine locations, co-located to leverage other investments.

- AmeriFlux sites provide the capacity to measure CO_2 fluxes in a variety of natural and managed ecosystems. A subset of sites would be equipped with advanced instrumentation for measuring fluxes of other important trace gases, such as nitrous oxide, methane, and ozone, for which fast, *in situ* sensors are only now becoming available or could be developed. These capabilities would lead to a better understanding of how such fluxes vary across environments, are affected by management decisions and natural disturbance, and contribute to climate.
- ARM Climate Research Facility observatories provide regional process-level information on atmospheric motions and hydrological processes. These observatories make a variety of measurements listed in detail on the ARM website (www.arm.gov). ARM facility enhancements would enable the program to operate in concert with advanced data assimilation systems to characterize regional aerosol, cloud, and precipitation processes and atmospheric dynamics in terrestrial and marine systems, information needed for developing and validating new climate models and biosphere feedbacks. An enhanced ARM facility will deliver a comprehensive suite of measurements within the three-dimensional volume of a climate model grid box (several to a hundred kilometers in size) or within the inner grid of a mesoscale model. Although a long-time ARM goal, this has not been fully realized at all sites, and thus improvements are needed.

Develop an Instrument Incubator Program. For the IFL concept to evolve as scientific understanding and modeling capabilities improve, advanced instrumentation must be deployed to provide observations that can help address the questions of coming decades. Several instrumentation needs were identified in the Long-Term Vision report (see list in Appendix 2, p. 35), and there are opportunities to expand existing Department of Energy (DOE) instrumentation platforms and facilities. Instrumentation improvement will become increasingly critical as scientific knowledge advances and greater measurement sophistication becomes essential. Foreseeing the specific instrumentation improvements that will be needed is difficult. However, advanced *in situ* and remote sensors with improved sensitivity and wavelength diversity undoubtedly will be needed, as will more-compact and -lightweight instruments with energy requirements suitable for field deployment in the subsurface, on towers, or on unmanned aerial vehicles. Additionally, better *in situ* sensing of clouds, precipitation, and aerosols will be required to provide increasingly detailed validation of emerging algorithms. To address these needs, an Instrument Incubator program should be started to provide sustained funding for developing the necessary instrumentation and ensuring its implementation. This likely would be a distributed program that draws on the shared expertise within BER user facilities, DOE national laboratories, and associated academic laboratories. A key component of the Instrument Incubator program must be the close interaction between those who design and build the instruments and those who use them in the field or laboratory.

3 Biosystems Frontier Network

Developing Tools to Predict Biological Response to Environmental Change

Microbes and plants in surface and subsurface environments catalyze a wide range of biogeochemical processes pivotal to Department of Energy (DOE) missions in energy and environment. These processes include the cycling of carbon, nutrients, energy, and water; capture and conversion of light and chemical energy; and production and consumption of greenhouse gases. Microbial and plant communities are highly sensitive and reactive to environmental stresses, and their responses in turn impact the environment. For example, water stress and temperature shifts likely associated with climate change will affect production of food crops and biofuel feedstocks. Despite the importance of biological responses to environmental conditions, the ability to measure and predict the physiological dynamics of these complex systems remains extremely limited.

The Biological and Environmental Research Advisory Committee's (BERAC) Long-Term Vision report (BERAC 2010) concludes that an ultimate goal in the biological sciences is attaining a deep and profound understanding of biological functions in complex systems so that living cells, cell communities, plant tissues, whole plants, and ecosystems can be reliably modeled. Achieving this level of understanding would enable prediction of and potential influence over how these biological entities respond to and affect diverse, dynamic, and evolving environments. Underpinning such predictive biology is, as identified in the Long-Term Vision report, the essential need to develop a complete "parts list" for a living cell, a community, and eventually an ecosystem. Improved quantitative measurement and imaging capabilities also are required for enhanced resolution, comprehension, dynamics, and real-time analyses of plant and microbe functional phenotypes, or "phenomics." These new capabilities must include real-time, nondestructive methods that enable comprehensive, cost-effective, and *in situ* analyses of hundreds and thousands of organisms across expansive temporal (nanosecond to years) and spatial (subnanometer to kilometer) scales and in diverse environments.

Advancing Plant and Microbial Phenomics with the Biosystems Frontier Network

BERAC proposes the creation of a Biosystems Frontier Network that would develop, test, and deploy the best approaches to study complex biological systems (see Fig. 3.1. Biosystems Frontier Network, p. 16). This network would generate novel tools and resources required to reveal, quantify, and define diverse plant and microbial phenomic parameters that represent dynamic physiological states in diverse environments and in the context of organismal and environmental interactions and feedbacks. Data generated through the network will be captured in integrated databases and analyzed in-depth to produce new knowledge and hypotheses about these systems.

The highly valued measurement and imaging capabilities envisioned for the Biosystems Frontier Network would significantly advance environmental and biogeochemical science by:

- Enabling much broader identification of biochemical and cellular characteristics and

microenvironmental conditions responsible for cell-to-cell variations observed in clonal and heterogeneous populations.

- Quantifying processes that occur at the interface between organisms and their micro- and macroenvironments.
- Assessing the spatial relationships, physical connections, and chemical exchanges that facilitate the flow of information and materials among proteins, cells, pore fluids, and minerals.
- Dynamically monitoring selected individual cells within a population.
- Performing targeted, real-time, and ultrafast measurements of metabolism and protein function for selected organisms. Such measurements would encompass a wide range of environmental conditions and fluxes that naturally occur in soils, subsurface environments, and sediments, as well as under perturbed conditions including temperature shocks, carbon dioxide (CO_2) elevation, and water stress.

Some of the general principles of this network can be found in the *DOE Genomics:GTL Roadmap* (U.S. DOE 2005). Although this roadmap expressed the concept of integrated resources to address biological systems, it envisioned them as specific, centralized user facilities. The Biosystems Frontier Network would be a distributed resource, encompassing existing Office of Biological and Environmental Research (BER) user facilities and also leveraging the collective expertise and capabilities of DOE national laboratories and academic collaborators. Because integration, cooperation, and collaboration are central to this resource, the Biosystems Frontier Network is fully

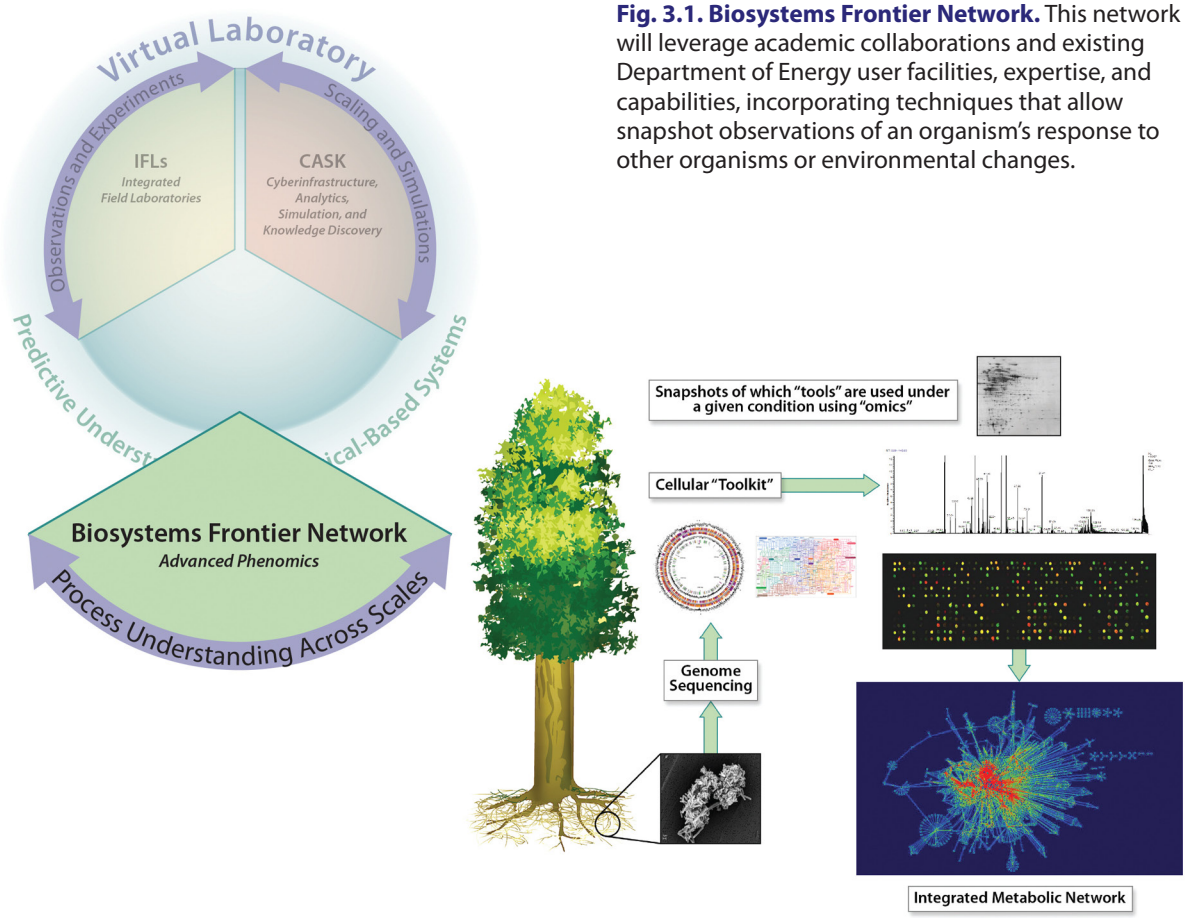


Fig. 3.1. Biosystems Frontier Network. This network will leverage academic collaborations and existing Department of Energy user facilities, expertise, and capabilities, incorporating techniques that allow snapshot observations of an organism's response to other organisms or environmental changes.

consistent with the goals of mission-inspired fundamental research outlined for the DOE Genomic Science program (genomicscience.energy.gov/pubs/GenomicScience-brochure.pdf).

Network to Complement and Facilitate Linkages to Other BER Assets

The Biosystems Frontier Network would directly link to integrated field laboratories (IFLs) by using subsets of materials and organisms for focused mechanistic studies. The environmental conditions of an IFL could be used to drive experimental manipulations in the Biosystems Frontier Network in micro- and mesocosms. This approach would allow observations in an IFL to be investigated at different scales and complexity using various biological and environmental tools developed by or available through the network.

Technology development is a key goal of BER user facilities and is among the common capabilities of national laboratories and the BER-supported Bioenergy Research Centers (genomicscience.energy.gov/centers/). Hence, many of the core competencies required for the Biosystems Frontier Network probably already exist within the BER science family. Yet a better mechanism clearly is needed to mold these capabilities into a coherent and efficient network that can achieve the level of coordination and integration required to meet complex biological challenges (see Sidebar 4, Addressing the Needs of 21st Century New Biology, p. 18). For example, advances in high-throughput gene and transcript sequencing have transformed the knowledge of microbial and plant genotypes and inferred function. However, translating this knowledge into predictions of dynamic function and physiological responses in the environment is still in its infancy and remains a significant challenge. New capabilities are needed that can provide a context for genomic measurements, leading to a higher-order understanding of biological complexity. These technologies include those that already exist at BER facilities but require much greater throughput and other capabilities that can be linked to biochemical and cellular function in the context of complex and diverse physical, chemical, and biological environments. An example is quantitative measurements of biomolecules (e.g., proteins and metabolites) at the Environmental Molecular Sciences Laboratory (EMSL) that can be enhanced by tools examining protein structure and activity, metabolic fluxes, physiology, and biogeochemistry. An additional key challenge is to provide functional information on the vast number of open reading frames coding for proteins of unknown function, whose numbers are exponentially rising because of the ease of whole-genome sequencing (see Sidebar 5, Annotating the Unknown, p. 19).

Another way the Biosystems Frontier Network would leverage BER resources is by complementing the methodologies and goals of the DOE Joint Genome Institute (JGI) as it seeks to become a next-generation genome center. While DOE JGI focuses on genomics, other components of the Biosystems Frontier Network can collaborate on proteomic measurements, protein function and activity studies, and metabolomic analyses and characterization. Because genomics informs proteomics and metabolomics, new correlative and causal relationships among these cellular components will emerge within the framework of the Biosystems Frontier Network as it combines the strengths of multiple facilities and fosters powerful synergies among them (see Sidebar 4, p. 18). The network will benefit from DOE JGI expertise in massive-scale sample preparation, single-cell technologies, and transcriptomic analyses. For example, single-cell isolation will enable determinations of single-cell metabolic and protein activity profiles. The network will enable researchers to perform and analyze hundreds of thousands of such measurements, providing data essential for modeling complex systems. Furthermore, high-throughput microbial phenotyping of mutagenized populations will complement the *in situ* phenomics of the Biosystems Frontier Network.

Sidebar 4

Addressing the Needs of 21st Century New Biology

A complex system is defined by a large number of interacting heterogeneous components that exhibit emergent properties, which often cannot be defined solely by measuring individual components in isolation. Hence, efficiently applying modern methods to achieve a true understanding of complexity requires interdisciplinary and well-integrated and -managed approaches. A common theme of the science supported by the Office of Biological and Environmental Research (BER) is the need to apply and improve approaches for understanding and predicting the behavior of complex systems. This necessitates a new BER structure emphasizing high-order collaboration. To specifically address such a need, this report calls for the formation of a Biosystems Frontier Network that combines capabilities in genomics, proteomics, metabolomics, fluxomics, functional and structural characterization, and synthetic biology and supports high-throughput application of each.

Many of these capabilities exist in BER-supported user facilities and Department of Energy (DOE) national laboratories, but more focused investment, stronger coordination, and better integration of distinctive and expert capabilities are needed to provide the required efficiency gains and necessary economy of scale. This coordination also should include integration with the DOE Systems Biology Knowledgebase (see Sidebar 6, p. 24) and the computational resources outlined for the cyberinfrastructure, analytics, simulation, and knowledge discovery (CASK) effort (see Chapter 4, p. 23).

An example of BER resources that would benefit from greater integration involves two of the program's major user facilities: the DOE Joint Genome Institute (JGI) and the Environmental Molecular Sciences Laboratory (EMSL). JGI focuses on sequence-based measurements, and EMSL uses integrated experimental (e.g., mass spectrometry, proteomics, nuclear magnetic resonance spectroscopy, microscopy, and imaging) and computational (e.g., NWChem software and advanced modeling) capabilities for molecular science research. These individual capabilities are very complementary and essential for addressing BER-relevant challenges in systems science. However, as currently structured, the facilities work too much in isolation, supporting independent community projects and lacking an incentive system to encourage collaboration. The same can be said in general about other BER-supported user facilities and many of the projects conducted at national laboratories. Greater collaboration among these resources could accelerate biology's role in providing practical solutions to many of the complex challenges facing the United States and world in the 21st century. This concept is the centerpiece of a 2009 National Research Council report (NRC 2009) that advocates a "New Biology" approach characterized by greater integration and collaboration within the biological sciences and the application of interdisciplinary approaches.

Although the current system clearly generates excellent science, it does not achieve its full potential. This report thus calls for the development of an incentive system to encourage collaboration among BER user facilities, DOE national laboratories, and the academic research community to align scientific capabilities with the needs of systems science.

The network also will capitalize on ongoing research and technology investigations. In 2010, for example, BER initiated a major effort in systems biology research focused on the role of microbial communities in carbon cycling. Emphasis is on regulatory and metabolic networks; metatranscriptomic-, metaproteomic-, and genomic-enabled approaches; and development of methods for imaging and analyzing microbially mediated carbon cycling processes. The 15 awarded projects encompass a wide variety

Sidebar 5

Annotating the Unknown

Modern high-throughput sequencing platforms enable scientists to generate in one day the same amount of sequence information they used to produce in a year or more. This expansion of sequencing capacity has made metagenomic and metatranscriptomic analyses of environmental samples accessible. Current efforts focus on transitioning from gene lists to an inference of phenotypes or "phenomics." However, meaningful extrapolations require unambiguous annotation of the genes present and expressed in a sample. Accurate annotation is a major limitation because the cataloging of a gene product's function currently is based almost solely on simple sequence similarity to a functionally characterized homolog, which itself may be flawed or incomplete. Given that 40% or more of the called open reading frames in a new genome sequence can be "unknowns," the unannotated portion of genome databases is growing exponentially with the increasing ease of DNA sequencing. Hence, accurate annotation—most importantly of open reading frames of unknown function—is a key challenge in genomic science.

Considering the magnitude of the problem, computational approaches probably will continue to provide the first line of analysis in trying to assign some function to each open reading frame. Given this reality, greater efforts are needed to improve current methods so that they include more than simple sequence comparisons and build upon new knowledge as it accumulates. Approaches that integrate methods based on dissimilar datasets (e.g., sequence comparison, protein-protein interaction, and structural similarity) are especially encouraged. Beyond this, mutant studies also can be useful because even though they do not provide a detailed biochemical description of function, they can reveal information on the phenotypes associated with a specific gene product, thus giving some clues as to function. Particularly informative are examples in yeast in which the use of double mutant approaches has revealed phenotypes associated with genes where single mutations showed no phenotypic changes. Conditional lethal mutations can be used to characterize cells in which mutations might lead to cell death.

Computational and mutant studies can provide the means to analyze the coding component of an entire genome but are not totally satisfactory since they do not provide a detailed description of a gene product's biochemical activity. For this, only a thorough biochemical analysis will suffice. Automated modern methods are available for characterizing proteins and should be applied to subsets of particular interest (e.g., reference enzymes and novel activities discovered in environmental samples). Because the number of enzymology experts has dwindled over the last decade, a facility specializing in biochemical and enzymological research is needed to assist with meticulous biochemical analyses. The scientific community can turn to such a facility for intensive short-term training, help with assay development, and interpretation of resulting data. A library of reference enzymes, their sequences, and structures would be developed. The facility also would provide high-throughput screening of enzymatic conditions, substrates, and inhibitors and become a resource of identified metabolites and chemical analogs available for enzyme screening. This facility could serve as a coordinating body so that a tiered series of approaches—starting from computational to mutagenesis, to high-throughput biochemical analysis, and ultimately to detailed biochemical characterization—could be applied to tackle the challenge of functional genome annotation.

of environments, ranging from individual consortia to prairies, wetlands, high-latitude permafrost, and marine areas.

Specific Biosystems Frontier Network Recommendations

Develop Comprehensive, Real-Time, and Nondestructive Methods to Quantify Stress-Specific Proteomic and Metabolomic Signatures for Microbial and Plant Communities in Diverse and Dynamic Hydrological and Biogeochemical Environments. Such signatures will report the impact of climate and environmental changes on the surrounding biosphere in near-real time, aiding detection of environmental stress and thus the ability to buffer its effects and responses through protective measures. Stresses such as temperature shocks can induce complex evolutionarily conserved responses in living organisms. However, comprehensive knowledge of diverse stress signatures in microbes and plants are poorly defined, including the limitations of coping with various stresses by individual members and the community as a whole. Proteins involved in the regulation of heat shock responses influence the speed of evolution, suggesting that stress responses can profoundly affect organismal fitness, productivity, and evolution. Specific protein activity and metabolite signatures arising from stress exposure could be used as indicators and warning systems of biocommunity health and climate change effects. For example, interactions between plants and an environment under stress can be strongly influenced by the presence of plant pathogens and their interplay with other mutualistic and commensal microbial species. These interactions may significantly affect the phenotypic information in a plant. Understanding this complex interplay between plants and microbial communities under various stresses is critical for modeling and predicting the behavior of communities on a macroscopic level, thus linking molecular properties and phenomics across time and length scales.

Experimental expertise within the Biosystems Frontier Network would enable comprehensive analysis of stress-specific proteomic, metabolomic, lipidomic, and glycomic signatures across diverse environments and communities. This network could comprehensively address various challenges, such as coupled carbon cycling and biofuel processes from the molecular to global scale, and could help measure transformations between different niches and environments. Critical questions could be examined concerning, for example, how food crops and biofuel feedstocks will respond to temperature and water stress and how genotype changes will affect plant and community phenotypes. Importantly, the information gathered and analyzed by network efforts will determine whether responses to environmental or genetic changes can be predicted.

Develop *In Situ*, Ultrafast, and Nondestructive Measurement and Analysis Techniques to Improve the Detection, Identification, and Resolution of Cellular Molecules, Metals, and Ions in Complex Environmental Samples. Although current methodologies can detect cellular proteins, protein activities, metabolites, metals, and ions, there are crippling limits of resolution, sensitivity, measurement throughput, integration, scale, and comprehensiveness. [These challenges are described more fully in two reports: *New Frontiers in Characterizing Biological Systems* (U.S. DOE 2009) and Long-Term Vision (BERAC 2010).] Enhanced detection, identification, and resolution of metabolites, proteins, protein function, metals, and ions are required. For example, although proteomic capabilities have advanced greatly over the past decade, there is a general inability to characterize the true complexity of the proteome. Present approaches are often blind to protein post-translational modifications that, in combination with alternative RNA splicing and protease activities, can give rise to millions of different *proteoforms* that potentially have quite different biological roles and activities. A further complication is the fact that enzymatic activity does not necessarily correlate with protein abundance, not only because of a lack of information on

most proteoforms, but also because their cellular localization and other structural details often are not known. Furthermore, current capabilities typically require protein extraction and thus lose spatial and temporal resolution.

Develop Advanced Bioimaging Platforms and Tools Featuring Multimodal Technologies with Ultrafast Temporal and Subnanometer Spatial Resolutions. Current bioimaging technologies are limited by a focus on a single imaging modality (e.g., fluorescence microscopy, three-dimensional electron microscopy, x-ray imaging and tomography, or mass spectroscopic metabolite imaging). These single approaches are in contrast to an integrated application of multiple imaging modalities that generate different types of data and complement and augment each other over spatial and temporal resolutions. In conjunction with the cyberinfrastructure, analytics, simulation, and knowledge discovery (CASK) component, the Biosystems Frontier Network will develop bioimaging platforms that (1) feature multifunctional probes and labeling chemistries providing contrasts across multiple imaging modalities; (2) advance image analysis, registration, and feature measurement algorithms to quantify images; (3) integrate and relate information across modalities; and (4) provide visualization and interaction systems for interpreting knowledge embedded in the data. These needed bioimaging developments and tools would complement and extend existing and developing EMSL capabilities applicable to both the biological and environmental sciences (see Sidebar 1, BER User Facilities Provide Core Capabilities for the Virtual Laboratory, p. 5). Tools developed and deployed by the Biosystems Frontier Network also would enable the following needed technical advancements [see p. 39 in *New Frontiers in Characterizing Biological Systems* (U.S. DOE 2009)].

- Mass spectrometry approaches that provide comprehensive proteomic, metabolomic, lipidomic, and glycomic information and enable broad spatiotemporal measurements.
- Single-cell metabolomic and proteomic measurements.
- Mass spectrometry for metabolomic and metaproteomic measurements applicable to field studies.
- Imaging approaches that provide both spatial and chemical information using synchrotron radiation on both the single-cell and community level.
- Rapid, multiresolution, multidimensional imaging technologies, including x-ray imaging.
- Super-resolution optical spectroscopy at the nanometer scale.
- Electrochemical imaging at the cellular level.
- Novel isotope technologies including subcellular tracer studies.
- Improved isolation and identification of metabolites from natural samples.
- X-ray crystallography to study structure-function relationships.
- Nuclear magnetic resonance to distinguish subtle biological molecular variation.
- Atomic force microscopy at the molecular level for soft biology.
- Higher-resolution electron microscopy for hydrated samples.
- Secondary ion mass spectrometry at the nanoscale.
- Nondestructive imaging and analytics for dynamics across long time scales.
- Synchrotron-based approaches including infrared x-ray fluorescence and tomography for *in situ* analyses.
- Studies of macromolecular assemblies using electron microscopy and synchrotron radiation.
- Hybrid methods to study complex systems, their interactions, and dynamics.

4 CASK—Cyberinfrastructure, Analytics, Simulation, and Knowledge Discovery

Tackling the Challenges of "Big Data"

Programs within the Department of Energy's (DOE) Office of Biological and Environmental Research (BER) are generating many valuable datasets and databases, including those associated with observatories for subsurface biogeochemistry, the AmeriFlux Network, and the Atmospheric Radiation Measurement (ARM) Climate Research Facility. Other sources of important data are BER-supported climate models and, most recently, the DOE Systems Biology Knowledgebase (KBase; see Sidebar 6, p. 24), which seeks to integrate genomics and systems biology data for microbes, plants, and microbial communities. However, these data collections are not being used to their full potential in advancing BER science for several reasons. They consist of disparate data streams sampling diverse spatial and temporal scales and cannot easily be related to other measurements that sample different Earth system properties at different scales. In some cases, the datasets can be prohibitively voluminous and unwieldy, creating computational and data management challenges when trying to extract physical principles and infer the physical processes the datasets implicitly document. The challenge of gaining greater insights from the wealth of available data is common across biological and environmental science. Indeed, Big Data management and analysis are now recognized as general problems that transcend all areas of science and engineering. In response, the White House announced in March 2012 a Big Data research and development initiative to address these issues (www.whitehouse.gov/blog/2012/03/29/big-data-big-deal/). Within BER, KBase was developed to tackle one component of the Big Data problem. Additional capabilities are greatly needed for facilitating transitions from multiscale, multitype datasets to information systems to web-enabled analytics, simulation, and knowledge discovery and then eventually to predictions of multiscale environmental and energy systems.

Revealing New Understanding of Big Data Using CASK

The Biological and Environmental Research Advisory Committee recommends developing a cyberinfrastructure, analytics, simulation, and knowledge discovery (CASK) component of the BER Virtual Laboratory. CASK would serve as a data and computational framework for handling large and diverse datasets, numerically representing key processes across vast space and time scales, and discovering knowledge about biological system behavior under a changing climate (see Fig. 4.1. CASK—Cyberinfrastructure, Analytics, Simulation, and Knowledge Discovery, p. 25). Each of these steps requires significant investment in software, hardware, and technology development closely linked with process science and natural-system expertise. For example, scientific and modeling applications that use BER data require information systems that synthesize multiple datasets of the same parameter; develop gridded data; provide essential metadata; and offer information on the accuracy, precision, and uncertainties of the different levels of processed data. The goal is not to merely store and curate information but to attain new knowledge and the whole-systems understanding that is a prerequisite for developing a new class of sustainable solutions to environmental and energy challenges.

Sidebar 6

Department of Energy Systems Biology Knowledgebase

The Department of Energy Systems Biology Knowledgebase (KBase, **kbase.us**) is a new community resource for predictive biology. It integrates a wide spectrum of data types across the microbial, plant, and microbial community domains and ties the data into an extensive set of powerful computational tools. These tools can analyze and simulate data to predict biological behavior, generate and test hypotheses, design new biological functions, and propose new experiments. The overarching objective is to provide a solid platform that supports predictive biology in a framework that does not require users to learn separate systems to formulate and answer questions spanning a variety of topics in systems biology research.

KBase is an open, extensible system that lowers barriers for using sophisticated algorithms and complex datasets to infer predictive models of biomolecular, organismal, and community function. The initial public offering of KBase tools and services is set for late February 2013 and will provide a programmatic environment featuring core biological analysis and modeling functions, including tools for linking multiple bioinformatics tools and datasets.

As KBase matures, there will be an increasing focus on enabling all biologists to work with this powerful system to essentially "build papers in place." KBase designers are developing an advanced web-based environment that will allow nonprogrammers to analyze complex datasets with effective computational workflows and share their reasoning and conclusions with the scientific community. Version 2.0 of KBase, planned for release in late 2013, will provide this advanced environment, called the narrative interface.

KBase is designed to increase the transparency, reproducibility, and collaborative nature of science by:

- Enabling scientists to integrate their algorithms and data into the system.
- Promoting sharing of both workflows and ideas in a socially enabled environment.
- Creating a new model for publication and review of computationally supported thought processes so that new biological conclusions and hypotheses can be achieved through such sharing.

KBase offers various types of training and collaboration with users and developers, including tutorials based on current KBase tools and interfaces, boot camps focused on teaching community developers how to create computational tools in KBase, and webinars and workshops targeting particular scientific problems. These training opportunities can be found at **www.kbase.us/about/training**/.

KBase is designed as a community-driven resource, and the project's outreach team serves as an advocate for the community in the KBase development process. Engaging in a dialogue with this team provides information that will help KBase set priorities for development, such as adding new analysis capabilities, data types, and datasets and improving the interfaces through which users can interact with KBase.

Coupled with the integrated field laboratories (IFLs) and Biosystems Frontier Network, the CASK component of the Virtual Laboratory would provide a critical vehicle for hypothesis and theory testing, data integration, and predictions of the influence of smaller-scale processes on the functioning of the larger system (i.e., ecosystem) and vice versa. CASK should advance the following key characteristics:

- **Database Linkages.** Because a single database cannot serve all the needs of the BER community, strategies should be developed and implemented for federation and exchange of data collections among resources, including data repositories, databases, and knowledgebases. Furthermore, new approaches such as cloud computing may make seamless access to public data feasible for users and provide data and analysis tools in a scalable fashion (an approach used by KBase). A major challenge is determining how best to maintain and curate these databases over the long term.
- **Assimilation of Data and Knowledge into Models.** CASK would facilitate integration of heterogeneous data and knowledge derived from that data (e.g., improved parameterizations and model parameters) into models via intuitive interfaces that advance discovery and knowledge development. The goal is to accelerate the use of these data—available from the Biosystems

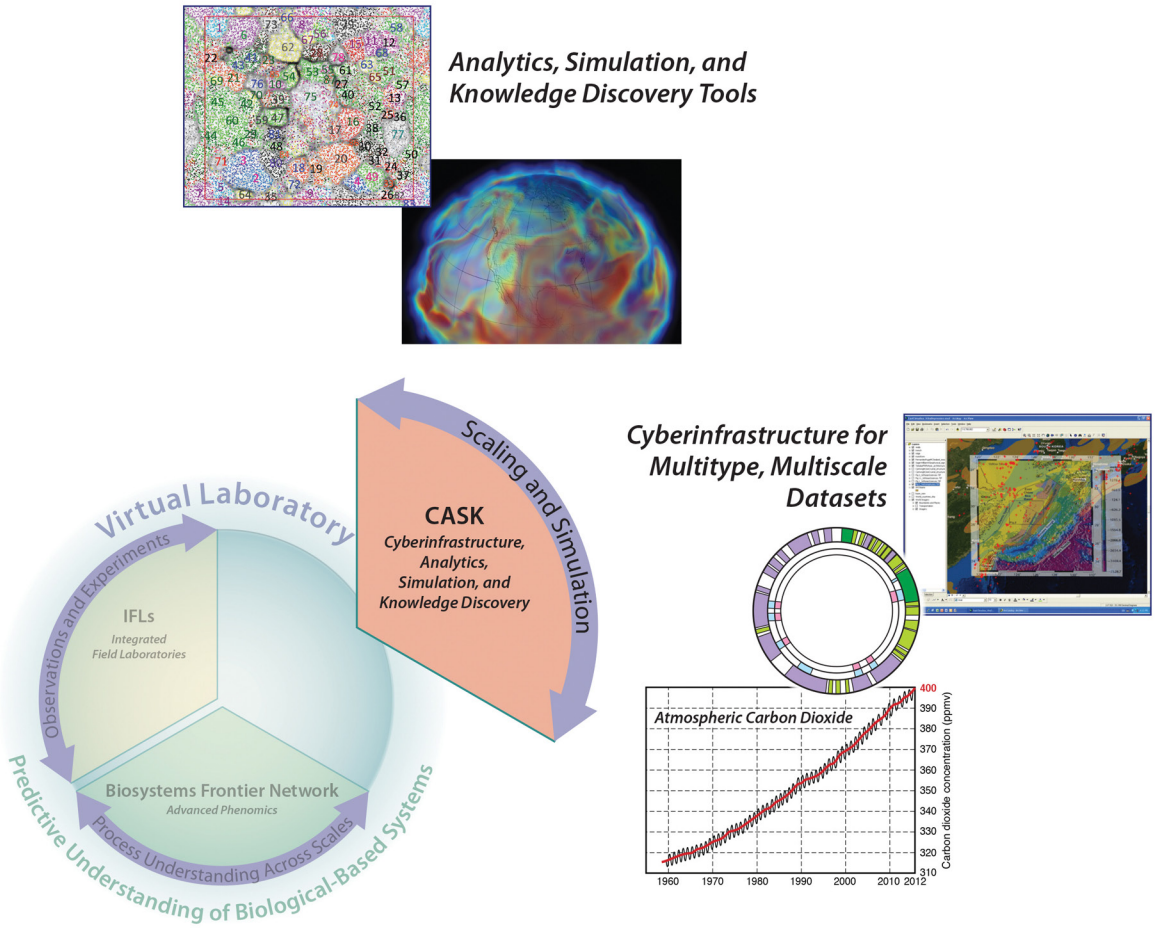


Fig. 4.1. CASK—Cyberinfrastructure, Analytics, Simulation, and Knowledge Discovery. The CASK component will amalgamate data from integrated field laboratories and the Biosystems Frontier Network, link databases, assimilate data and knowledge into models, advance multiscale modeling, and promote knowledge discovery.

Frontier Network, IFLs, climate models, and other resources—to provide a comprehensive analysis of biological and environmental systems, thereby advancing system understanding and model predictions and fidelity.

- **Multiscale and Advanced System Component Models.** CASK would both advance individual system component models and develop new approaches for bridging computation and natural phenomena representation across vast temporal and spatial scales, from molecular to global. The new approaches are expected to benefit from novel mathematical constructs and process-based theory and understanding. Once developed, the simulation capabilities could be used to identify the greatest sources of model uncertainty, critical thresholds and tipping points, and sensitivities in system response, all of which can, in turn, drive and prioritize process investigations and observations.
- **Knowledge Discovery.** A key CASK component is development of a knowledge discovery environment that provides innovative capabilities in distributed data discovery, visualization, and analysis (including uncertainty). This component will leverage existing BER investments, such as KBase.

Specific CASK Recommendations

Link Heterogeneous Databases. Many of the most interesting and important future applications of BER datasets will require integrating terrestrial subsurface and land surface, marine, atmospheric, and biological (e.g., organisms, genomic, and molecular) information. Linking this information will involve specifically designing measurement sites appropriate for the eventual integration of datasets that cross system boundaries. Critical measurements that may not appear interesting for specific communities but could be important for understanding processes occurring at the interfaces must be identified and given priority. Common metadata standards and data collection and management protocols need to be developed and adopted, and data should be organized geospatially using geoinformatics.

Develop Multiscale Simulation Frameworks and Data Assimilation Tools. These tools are needed to facilitate hypothesis testing, assimilate multiscale data, and assess many fundamental issues pivotal to sustainable environmental and energy strategies that involve processes and their couplings ranging from molecular and cellular levels to the ecosystem scale. A variety of simulation approaches are needed, depending on the question asked, the data available, and the scales of interest. Examples include (1) mechanistic, hierarchical simulation frameworks that can transfer information across a range of spatial and temporal scales; (2) stochastic, analytical, and phenomenological simulation capabilities that allow scientists to synthetically explore system responses to perturbation as a function of selected processes; and (3) system simulation models that incorporate both anthropogenic drivers and system responses. New mathematics and simulation methods may be considered for modeling the emergent properties of nonlinear biological and environmental systems influenced by interactions among smaller-scale system components and by larger-scale external phenomena and conditions. Advances are needed to assimilate heterogeneous (streaming) datasets into simulations and thus effectively use all the different types of data that are and will continue to become available for systems understanding. These are areas that can benefit from closer coordination and cooperation between BER and DOE's Office of Advanced Scientific Computing Research.

Although simulation frameworks may include components that vary from application to application, they likely will have common approaches, architectures, features, modules, and couplers that can be used to advance different aspects of multiscale biological and environmental system predictions.

Develop Advanced System Component Models. As part of the CASK effort, strengthening individual system component models and the linkages between different models is imperative. Examples

of needed improvements include incorporating cell function into reactive transport models, coupling subsurface and watershed biogeochemical simulators to land models, and advancing DOE climate models as outlined below.

First, only a few studies have attempted to incorporate mechanistic microbial function into subsurface biogeochemical reactive transport models. Examples include the use of *in silico* (Fang et al. 2011) to ecotrait-based methods (Bouskill et al. 2012). Significant efforts are needed to advance these methods to allow simulation of cellular, organismal, or community responses to environmental fluxes; their impacts on the field environment; and the effects of field-scale biogeochemical, hydrological, and atmospheric fluxes on microbial community functioning. Considerable challenges lie in (1) developing and testing such coupled models; (2) determining the level at which microbial function and functional groups can and should be represented in reactive transport models; (3) representing microbe-microbe, microbe-plant, and microbe-mineral interactions; and (4) determining how to parameterize microbial functioning in a tractable yet representative manner.

Second, capabilities for simulating water flows within a watershed have improved over the last decade in the hydrological community. However, the development of mechanistic reactive transport models that also can simulate biogeochemical reactions coupled to these flows up to the watershed scale is in the early stages. Moreover, the coupling of reactive watershed models with land surface models is a frontier area. Improved understanding and methodologies are greatly needed to represent hydrological and biogeochemical connectivity across scales in a manner that honors landscape characteristics, hydrological boundaries, lateral and vertical fluxes and transformations, and subsurface heterogeneity and processes that influence biogeochemical cycling.

Third, in collaboration with the National Center for Atmospheric Research, DOE's climate modeling efforts are focused on a community modeling framework characterized by models with different modularity and configurations. Climate modeling over the next 5 years should include efforts to decrease the grid spacing of global coupled models to ~4 km and assess the tradeoffs between ensemble size and horizontal resolution for the range of relevant model applications. Also needed is continued work to quantify uncertainties in model results and incorporate greater completeness and fidelity in physical processes using ARM Climate Research Facility data, with the goal of minimizing tunable parameters and tying the physics to first principles. These climate modeling improvements require research to design advanced and computationally efficient numerics and algorithms and model coupling to take advantage of high-performance computers with over 10 million cores. Additional investigations are needed to develop approaches for modeling three-dimensional land surface and cryosphere processes effectively at high resolution. Furthermore, these models need to take advantage of petascale to exascale computers and the associated technology and software for input/output and data management, storage, visualization, and workflows. Another important challenge will be improving model physics for all system components. Addressing these challenges will enable global, high-resolution Earth system models to assess climate sensitivity and explicitly simulate regional climate.

Extend and Link Knowledge Discovery Tools. Methods are needed for analyzing interaction networks, from molecular to global scales, to gain insight into environmental and climate effects on terrestrial parameters. These methods might include tools to measure and visualize feed-forward and feed-back mechanisms controlling nutrient and signal exchange among organisms within an ecosystem. Other questions could relate to understanding the impact of various stressors on interactions between cells and organisms or the role of aerosols in clouds. Such knowledge discovery is needed to increase the information content of individual databases so that the data are more effectively used for science applications and integrated with models. Although information is implicitly contained within

datasets, the gulf between data, information, and actionable knowledge is broad and can be spanned only by developing and implementing technologies such as relational databases embedded in process models of the system within which the data have been collected. Future process models need to explicitly include linkages to databases and knowledgebases that form the basis for those models and therefore can be updated, improved, or removed as scientific understanding advances or changes. In particular, scientific and modeling applications that use BER data require information systems that synthesize multiple datasets of the same parameter; develop gridded data; provide essential metadata; and offer information on the accuracy, precision, and uncertainty of different levels of processed data. A framework that explicitly recognizes the needs of data curation, integration, refinement, and abstraction is essential. Each modeling community needs high-quality data products at the right level of abstraction and the provenance necessary for curation. BER's infrastructure should include open data repositories, databases, and knowledgebases, the latter containing the highest level of curated data that can be built on by others. A key issue in this regard is development of controlled vocabulary ontologies and semantic-oriented search tools to provide a broadly useful index and searching capabilities for finding records according to their scientific meaning, rather than syntax.

5 Strategic Demonstrations and Workshops

Cross-cutting themes of the Long-Term Vision report (BERAC 2010) include complex systems science across scales, multidisciplinary research, and computing and mathematics. Within the Virtual Laboratory envisioned for the Office of Biological and Environmental Research (BER), each theme could be brought to bear on key energy and environmental questions. The laboratory also would provide an integrated context for addressing important uncertainties in four areas of BER science:

- **Biological systems**, by measuring and analyzing such systems, exploring ecosystem function and elemental cycling, and ultimately enabling predictive biology.
- **Climate research**, by developing higher-resolution models, improving basic knowledge about aerosols, and advancing understanding of important biological interactions and feedbacks.
- **Energy sustainability**, by analyzing approaches for organizing land use, water use and quality, and energy systems sustainably; characterizing potential Earth system drivers, feedbacks, and vulnerabilities to state changes; and developing unifying models to test the significance of global change issues.
- **Computing**, by establishing new paradigms for data management and computing necessitated by data-intensive science and by designing and building software solutions that provide researchers with better access to large, complex, and interrelated datasets.

Strategic Demonstrations to Develop, Integrate Laboratory Components

Seamless interaction among the three components of the Virtual Laboratory will be imperative to their success. The integrated field laboratories (IFLs) will use discoveries from the Biosystems Frontier Network to address crucial questions at the scales of molecules, organisms, and their physicochemical microenvironments. Such questions will aid in understanding the role each plays in the biogeochemical processes affecting energy and climate, from the bedrock to the atmosphere. The cyberinfrastructure, analytics, simulation, and knowledge discovery (CASK) effort will collect, organize, and integrate the huge datasets resulting from field observations and experiments and will provide new simulation approaches for creating predictive models.

Strategic demonstrations are recommended as a strategy to guide the development and integration of the full-scale Virtual Laboratory. These demonstrations are expected to identify both the strengths and pitfalls associated with a fully deployed Virtual Laboratory and to initiate the design of a roadmap for laboratory implementation. BER's Next-Generation Ecosystem Experiments in the Arctic or Tropics could be expanded to meet the need for an initial strategic demonstration, while different subsurface biogeochemistry, surface vegetation, hydrology, and atmospheric processes could provide valuable contrasts for refining the roadmap.

Questions that the strategic demonstrations need to address should be identified and prioritized at a multidisciplinary planning workshop but potentially could include the following:

- How do microbial communities interact in different soils and subsurface environments to produce trace gases important to the atmosphere?
- What is the role of vegetation in producing and attenuating aerosols that affect cloud formation and atmospheric reflectance?
- How do the spatial and temporal variabilities of soil, water, and microbial taxa affect biogeochemical cycles, plant community structure, and carbon dioxide exchange between the plant canopy and the atmosphere?
- What role do extreme climate events play in structuring plant, microbial, and biogeochemical responses to environmental stress?
- How can the complex feedbacks among subsurface, surface, and atmospheric influences best be modeled to effectively simulate ecosystem carbon and nitrogen exchange?

Community Workshops: Steps Toward Laboratory Development

To facilitate development of Virtual Laboratory components and outline a path toward laboratory implementation, the Biological and Environmental Research Advisory Committee recommends BER convene four workshops:

- **Environmental Observatories Workshop** to identify and prioritize opportunities for expanding and integrating field laboratories, both vertically (from the bedrock to the atmosphere) and geographically (across key regions). This workshop should discuss ways to leverage BER and other community observatory investments.
- **Biosystems Frontier Network Workshop** to explore the development of a distributed research and instrumentation resource built upon the core expertise of existing BER user facilities (e.g., DOE's Joint Genome Institute and Environmental Molecular Sciences Laboratory) and augmented by an efficient network of collaborating national and academic laboratories. Such a resource would enable quantitative observations and *in situ* measurements of biomolecules that can be linked to function in the context of complex and diverse physical, chemical, and biological environments.
- **Multiscale Data Simulation and Assimilation Workshop** to create a detailed, prioritized roadmap for developing multitype, multiscale assimilation, simulation, and knowledge discovery capabilities needed to address cross-cutting BER grand challenges. This workshop would call on diverse experts from the scientific community to consider the technical, mathematical, theoretical, and computational challenges associated with data assimilation and strategies for fully leveraging BER's myriad databases and knowledgebases to gain a predictive understanding of multiscale phenomena.
- **Implementation Workshop** to synthesize and prioritize recommendations resulting from the three previous, component-specific workshops and to develop a roadmap for creating the BER Virtual Laboratory.

6 Checklist of Recommendations for Virtual Laboratory

This document describes many research needs and recommendations associated with the components, strategic demonstrations, and workshops proposed for the Department of Energy (DOE) Office of Biological and Environmental Research (BER) Virtual Laboratory. Summarized in this chapter are specific recommendations for the three laboratory components: integrated field laboratories (IFLs); Biosystems Frontier Network; and cyberinfrastructure, analytics, simulation, and knowledge discovery (CASK). Also described here are near-term, actionable recommendations needed to advance Virtual Laboratory development.

The Vision for Components

Integrated Field Laboratories

IFLs would measure elemental, energy, and water transfer across mineral, biological, and atmospheric interfaces, including all ecological and climate processes that play a role in geochemical cycling. The resulting observational data would provide objective functions and a framework to quantify how key mechanisms and processes at various scales couple to control essential system behaviors. Specific recommendations include:

- **Identify Opportunities for Leveraging and Extending BER Investments to Develop Vertically Integrated Laboratories.** New hypothesis-driven experimentation would be incorporated into existing and planned field sites, with the goal of achieving a network of horizontally and vertically linked laboratories focused on integrating research results from the Biosystems Frontier Network and CASK. These laboratories would address process-level questions in contaminant migration, global climate change, bioenergy development, and land use.
- **Strategically Identify Geographically Dispersed Sites with the Necessary Subsurface, Land Surface, and Atmospheric Components.** With properties essential for achieving the objectives outlined above, these sites would include and extend BER investments in existing observatories associated with subsurface biogeochemistry, the Atmospheric Radiation Measurement Climate Research Facility, and AmeriFlux Network, as well as other resources within the scientific community.
- **Develop an Instrument Incubator Program.** Such a program would address the need for new sensing and analytical capabilities required to understand water and chemical fluxes across physical, chemical, and biological interfaces and boundaries. Particular emphasis should be placed on developing dependable instruments that can be used in the field under rigorous environmental conditions.

Biosystems Frontier Network

New capabilities are needed to enable quantitative and high-throughput measurements and observations of biomolecules that can be linked to function in the context of complex and diverse physical, chemical, and biological environments. Specific recommendations include:

- **Develop Comprehensive, Real-Time, and Nondestructive Methods to Quantify Stress-Specific Proteomic and Metabolomic Signatures for Microbial and Plant Communities in Diverse and Dynamic Hydrological and Biogeochemical Environments.** Such signatures will report the impact of climate and environmental changes on the surrounding biosphere in near-real time, aiding detection of environmental stress and thus the ability to

buffer its effects and responses through protective measures.

- **Develop *In Situ,* Ultrafast, and Nondestructive Measurement and Analysis Techniques to Improve the Detection, Identification, and Resolution of Cellular Molecules, Metals, and Ions.** Capabilities with enhanced sensitivity, measurement throughput, integration, scaling, and comprehensiveness are needed to analyze, for example, metabolites, proteins, lipids, glycans, and protein function.
- **Develop Advanced Bioimaging Platforms and Tools Featuring Multimodal Technologies with Ultrafast Temporal and Subnanometer Spatial Resolutions.** Integrated application of these multiple imaging modalities will allow their different types of data to complement and augment each other across space and time scales.

CASK

The CASK component of the Virtual Laboratory would serve as a cyberinfrastructure for hypothesis and theory testing, data integration, and predictions of the influence of smaller-scale processes on the functioning of the larger system and vice versa. Specific recommendations include:

- **Link Heterogeneous Databases.** Important future applications of BER datasets will require integrating diverse information, including terrestrial subsurface and land surface, marine, atmospheric, and biological data.
- **Develop Multiscale Simulation Frameworks and Data Assimilation Tools.** These tools are needed to facilitate hypothesis testing, assimilate multiscale information, and assess many fundamental issues pivotal to sustainable environmental and energy strategies that involve processes and their couplings ranging from molecular and cellular levels to the ecosystem scale.
- **Develop Advanced System Component Models.** Examples of needed improvements are incorporating cell function into reactive transport models, coupling watershed biogeochemical simulators to land models, and advancing DOE climate models.
- **Extend and Link Knowledge Discovery Tools.** Although information is implicitly contained within datasets, the gulf between data, information, and actionable knowledge is broad and can be spanned only by developing and implementing technologies such as relational databases and methods for analyzing interaction networks, from molecular to global scales.

Workshops, Strategic Demonstrations

To orchestrate development of the BER **Virtual Laboratory**, the Biological and Environmental Research Advisory Committee (BERAC) recommends strategic demonstrations and four workshops be initiated in the near term.

With a focus on developing Virtual Laboratory components, three workshops should address environmental observatories, the Biosystems Frontier Network, and multiscale simulation and data assimilation. The fourth workshop should synthesize outcomes from the previous three meetings and generate a plan to implement the laboratory.

After the workshop series, BERAC recommends initiating strategic demonstrations to advance and test one to three prototype designs for the Virtual Laboratory. The strategic demonstrations are envisioned to guide full-scale development and integration by formulating and testing a prototype in conjunction with selected field laboratories, CASK components, and Biosystems Frontier Network capabilities. These demonstrations also aim to identify the benefits and limitations of the Virtual Laboratory concept. Guided by hypothesis-driven scientific questions that require predictive understanding of multiscale environmental phenomena, the strategic demonstrations will explicitly define how the Virtual Laboratory will leverage and integrate BER and other community resources.

By conducting these complementary activities in a collaborative manner, emerging recommendations ultimately will lead to development of the BER Virtual Laboratory.

Appendix 1: Charge Letter

Department of Energy
Office of Science
Washington, DC 20585

Office of the Director

September 14, 2011

Dr. Gary Stacey
Associate Director, National Soybean Biotechnology Center
Department of Microbiology and Molecular Immunology
University of Missouri
271 E Christopher S. Bond Life Sciences Center
Columbia, MO 65211

Dear Dr. Stacey:

In December 2010, the Biological and Environmental Research Advisory Committee (BERAC) prepared a report, "Grand Challenges for Biological and Environmental Research: A Long-Term Vision," that laid out grand research challenges for BER in biological systems, climate, energy sustainability, computing, and education and workforce training that can put society on a path to achieve the scientific evidence and predictive understanding needed to inform decision making and planning to address future energy needs, climate change, water availability, and land use. Two key goals were to: (1) describe how BER should be positioned to address those challenges; and (2) determine the new and innovative tools needed to advance BER science.

A recognized strength of the Office of Science, and BER is no exception, is the development of tools and technologies that enable science - from synchrotrons to genomic sequencing to nanoscience research centers. The BERAC report identifies technology needs that will be important for BER to achieve the scientific grand challenges outlined in the report. These ranged from the development of new observational technologies for biological systems, climate model integration, and energy sustainability, to the application of advanced computational and analytical capabilities to characterize network interactions. I am now charging BERAC to:

- Expand on the development and use of new tools that were only briefly mentioned in the "Long Term Vision" report;
- Identify the development and use of new tools and their linkage to existing or new user facilities;
- Identify linkages between new tools and existing resources, new resources and to diverse scales of time and space;
- Expand on the concepts of virtual laboratories and collaborative tools, including a discussion of how to facilitate these concepts and interactions.

I would like to receive a progress report on this charge at the early spring 2012 BERAC meeting and a final report at the fall 2012 BERAC meeting. I look forward to what should be a stimulating and useful report. Many thanks for your contributions to this important effort.

Sincerely,

W. F. Brinkman
Director, Office of Science

cc: Sharlene Weatherwax
David Thomassen

Appendix 1: Charge Letter

Appendix 2: Examples of Needed Technologies

[Note: This list was extracted from the Long-Term Vision document (BERAC 2010) and the *Complex Systems Science for Subsurface Fate and Transport* report (U.S. DOE 2010). It contains several similar needs identified in each report.]

1. Collaborative tools and systems using hyperspectral satellite imaging to quantify plant species and physiological function at the landscape level.
2. *In situ* methods to quantify how microbial communities respond to dynamic changes in environmental conditions.
3. Identification of model organisms for relevant environmental process understanding.
4. Analytical and computational methods to characterize network interactions in microbial communities (including microbe-microbe and plant-microbe interactions).
5. Reproducible methods for characterizing subsurface microbial processes, metabolites, and genomic properties (e.g., metaomics and infochemicals) and for establishing mass balances.
6. Methods to accurately measure and predict soil moisture and groundwater on seasonal and longer time scales.
7. *In situ* and *in vivo* methods for sampling microbes and biofilms without perturbing genetic expression and physiological function.
8. New approaches to address the issue of microbial nonculturability in natural systems.
9. Measurements and understanding of hydraulic redistribution within soils, especially within fractured media and as influenced by soil roots.
10. Databases on global soil characteristics (e.g., depth, texture, infiltration, permeability, and nutrient level) and above- and belowground vegetation properties.
11. Methods to quantify mechanisms and rates of organic carbon decomposition by microbes in permafrost soils (especially under variable and fluctuating oxygen and temperature states) and to assess their effective production of greenhouse gases at larger scales.
12. Methods to measure microbial community behavior at both subzero and above-zero temperatures and under a variety of soil moisture and geochemical conditions in permafrost.
13. Common *in situ* approaches to quantify *in situ* biogeochemical reaction rates across environments with different physical and geochemical characteristics and different microbial communities.
14. Detailed, long-term soil measurements in multiple landscapes across Earth.
15. Computing paradigms that can meet the enormous parallel processing and intensive analysis needs for biological, climate, and environmental data.
16. Software solutions that enable better access to increasingly large, complex, and interrelated datasets.
17. Framework for allowing process models to interact meaningfully across molecules to ecosystems to the whole Earth (i.e., see "Google Life" sidebar on p. 46 in BERAC 2010).

18. Analytical, visualization, and computational capabilities for assessing temporal and spatial heterogeneity in soil, plant, and subsurface systems (including organism-gene distribution; cells, proteins, and mineral-solid relationships; and "whole" macrosystem analysis).

19. Innovative use of low-complexity subsurface communities to aid in understanding more complex natural communities.

20. Plant and microbial genomic approaches for system-level understanding (e.g., the Department of Energy Joint Genome Institute).

21. Proxy or diagnostic signatures of critical system states or transformations using methods that are easier, less invasive, and cheaper to deploy or that have larger spatial and temporal extents than direct measurements.

22. Imaging and analytical approaches for subsurface microbes, solid phase properties, and solutes (e.g., SLAC National Accelerator Laboratory and Environmental Molecular Sciences Laboratory).

23. New methods for measuring key parameters such as pH, redox, and solutes within microbial communities and at the microbe–plant root level.

24. New experimental approaches and *in situ* sensors for characterizing and measuring biogeochemical dynamics at the microbe-mineral interface.

25. *In situ* and remote methods for quantifying and modeling feedbacks between microbe-mediated transformations in the subsurface and water flow.

26. New approaches and methods to quantify the influence of smaller-scale processes on higher-scale behavior (e.g., molecular to pore, pore to porous medium, porous medium to field, field to ecosystem, ecosystem to landscape, landscape to region, and region to globe).

27. Innovative strategies to interrogate large-scale system behavior and develop phenomenological models for understanding and prediction.

Appendix 3: Bibliography

Albrecht, B. A. 1989. "Aerosols, Cloud Microphysics, and Fractional Cloudiness," *Science* **245**(4923), 1227–30.

BERAC. 2010. *Grand Challenges for Biological and Environmental Research: A Long-Term Vision; Report from the Biological and Environmental Research Advisory Committee March 2010 Workshop*, DOE/SC-0135, BERAC Steering Committee on Grand Challenges for Biological and Environmental Research (genomicscience.energy.gov/program/beracltv.shtml).

Bouskill, N., et al. 2012. "Trait-Based Representation of Biological Nitrification: Model Development, Testing, and Predicted Community Composition," *Frontiers in Aquatic Microbiology* **3**, 364.

Fang, Y., et al. 2011. "Direct Coupling of a Genome-Scale Microbial *In Silico* Model and a Groundwater Reactive Transport Model," *Journal of Contaminant Hydrology* **122**(1–4), 96–103.

Jansson, J. 2011. "Towards 'Tera-Terra': Terabase Sequencing of Terrestrial Metagenomes," *Microbe* **6**(7), 309–15.

NRC. 2009. *A New Biology for the 21st Century*, National Research Council Committee on a New Biology for the 21st Century: Ensuring the United States Leads the Coming Biology Revolution (nap.edu/catalog.php?record_id=12764).

Stevens, B., and G. Feingold. 2009. "Untangling Aerosol Effects on Clouds and Precipitation in a Buffered System," *Nature* **461**, 607–13.

Twomey, S. 2007. "Pollution and the Planetary Albedo," *Atmospheric Environment* **41**, 120–25.

U.S. DOE. 2005. *Genomics:GTL Roadmap: Systems Biology for Energy and Environment*, DOE/SC-0090, U.S. Department of Energy Office of Science (genomicscience.energy.gov/roadmap/).

U.S. DOE. 2009. *New Frontiers in Characterizing Biological Systems: Report from the May 2009 Workshop*, DOE/SC-0121, U.S. Department of Energy Office of Science (genomicscience.energy.gov/characterization/).

U.S. DOE. 2010. *Complex Systems Science for Subsurface Fate and Transport: Report from the August 2009 Workshop*, DOE/SC-0123, U.S. Department of Energy Office of Science (doesbr.org/complexityreport/).

U.S. DOE. 2012. *DOE Joint Genome Institute Strategic Planning for the Genomic Sciences: Report from the May 30–31, 2012, Workshop*, DOE/SC-0152, U.S. Department of Energy Office of Science (genomicscience.energy.gov/userfacilities/jgi/futuredirections/).

Wrighton, K. C., et al. 2012. "Fermentation, Hydrogen, and Sulfur Metabolism in Multiple Uncultivated Bacterial Phyla," *Science* **337**(6102), 1661–65. DOI: 10.1126/science.1224041.

Acronyms

ARM	Atmospheric Radiation Measurement
BER	Office of Biological and Environmental Research
BERAC	Biological and Environmental Research Advisory Committee
CASK	cyberinfrastructure, analytics, simulation, and knowledge discovery
DOE	Department of Energy
EMSL	Environmental Molecular Sciences Laboratory
IFLs	integrated field laboratories
IFRC	integrated field research challenge
JGI	Joint Genome Institute
KBase	Systems Biology Knowledgebase
LTER	Long-Term Ecological Research program
NASA	National Aeronautics and Space Administration
NEON	National Ecological Observatory Network
NGEE	Next-Generation Ecosystem Experiment
NSF	National Science Foundation