Scientific User Research Facilities and Biological and Environmental Research: Review and Recommendations

A Report from the Biological and Environmental Research Advisory Committee

October 2018

Gary Stacey, Chair
Biological and Environmental Research Advisory Committee (BERAC)

Bruce A. Hungate, Chair
BERAC Subcommittee on Scientific User Research Facilities

U.S. Department of Energy

Prepared by the BERAC Subcommittee on Scientific User Research Facilities

Report available online at
https://science.energy.gov/ber/berac/

Contents

Chapter 1. Executive Summary

This report summarizes findings of a subcommittee of the Biological and Environmental Research Advisory Committee (BERAC) to the U.S. Department of Energy (DOE), assessing the role of User Facilities in supporting research funded by the Office of Biological and Environmental Research (BER) within DOE's Office of Science (see Appendix A. Charge Letter to BERAC from the DOE Office of Science, p. 84). Specifically, the subcommittee evaluated:

1. Optimal alignment of User Facilities to support the current BER research portfolio,
2. Optimal alignment of User Facilities to support future research needs identified in the 2017 Grand Challenges report,[1]
3. Development of additional User Facility capabilities, and
4. Opportunities to collaborate between User Facilities (internal to DOE and also external interagency partners).

The subcommittee also evaluated five primary topical areas within the scope of BER research: Biological Systems Science, Earth and Environmental Systems, Microbial to Earth System Pathways, Energy Sustainability and Resilience, and Computation and Data Analysis (see details below). User Facilities are those institutions, installations, and resources that are available to the scientific community to support research by providing access to state-of-the-art analytical tools, field experiments, and computational resources. Overall, BER's User Facilities—the Joint Genome Institute (JGI), Environmental Molecular Sciences Laboratory (EMSL), and Atmospheric Radiation Measurement (ARM) research facility—provide essential and unique support well aligned with much of the current BER research portfolio. Better integration between facilities, strategic investments in human resources and equipment, and better communication of the available capabilities to more members of the scientific community will improve how User Facilities are situated to address the 2017 Grand Challenges.

Emergent Themes

BER research addresses "complex biological and environmental processes that range from molecular to global scales over time horizons of nanoseconds to centuries and beyond." User Facilities are especially well positioned to contribute to science that scales from molecules to the planet, and future investments in facilities could help advance this science of scaling and quantitative synthesis. The strong theme of research across scales articulated in many of the recommendations (see Appendix B. List of Recommendations, p. 86) in this document (e.g., connecting "omics" to ecosystems, advocating efforts to develop principles for translating across scales of space and time, and establishing a User Facility based on modeling across scales) speaks to the potential of User Facilities to lead in this integration. User Facilities are well positioned to integrate research across the full purview of BER research, grounded in the mechanisms of molecular events that happen at rapid rates, with implications for organisms and ecosystems and up to global scales that are impacted for centuries and beyond. Needs for computational synthesis, training, and big data management were also strong themes that emerged as strategic points of focus for User Facility improvement.

---

[1] BERAC. 2017. *Grand Challenges for Biological and Environmental Research: Progress and Future Vision; A Report from the Biological and Environmental Research Advisory Committee*, DOE/SC-0190, BERAC Subcommittee on Grand Research Challenges for Biological and Environmental Research (science.energy.gov/~/media/ber/berac/pdf/Reports/ BERAC-2017-Grand-Challenges-Report.pdf).

Biological Systems Science (BSS)

The BER User Facilities support current BER research well, and this assessment matches the community survey of existing User Facility resources to support work in biological systems science (see Figure 1, p. 4, and Box 1, p. 5). Research goals and Grand Challenges within BSS are well supported by the existing network of User Facilities, especially JGI and EMSL, and also by User Facilities beyond BER. This is consistent with the assessment offered in the 2017 Grand Challenges (see Table ES-1 of that document). A number of specific recommendations emerged for BSS research where User Facilities can help address these Grand Challenges (see Chapter 2. Biological Systems Science, p. 7). Many of them capture strategic points of research focus, such as developing metabolic pathway databases based on experimentally annotated gene function, structural libraries for metabolites and enzymes, and stoichiometric and kinetic models of metabolism that integrate omic and isotopic data with metabolic flux analysis. Other recommendations are more specifically targeted to instrument acquisition and capacity development, such as developing new tools for engineering organisms, for porting biosynthetic pathways between organisms, for genome disruption, and for new cellular sensors to monitor metabolism. User Facilities should acquire new analytical capacity for measuring intracellular and interspecies fluxes of metabolites; for quantitative and standardized stable isotope probing; for DNA synthesis; for protein synthesis; and for microfluidics and nanotechnologies, cryo-electron microscopy (cryo-EM), "multiomics," and label-free imaging. Improvements in data management and computation were also identified, including tools for predicting gene function; a shared search platform across User Facilities; metadata standards; data storage and computation; and collaborative research calls for interdisciplinary work in biological systems science, mathematics, and computation. The need for interdisciplinary training was also identified. Finally, across the multiple User Facilities supporting BSS research, there was interest in the development of a Coordinated Network for Systems Biology (CNSB), a multisite User Facility network comprising existing User Facilities, designed to address large-scale and complex challenges in biological systems science.

Earth and Environmental Systems (EES)

EES is a key area of scientific focus within the current BER research portfolio. The survey of existing User Facilities conducted as part of this assessment (summarized in Figure 1, p. 4) shows that core BER User Facilities (ARM, JGI, and EMSL) are either well aligned with or have potential to align with research addressing some of the Grand Challenges in Earth and environmental systems science; facilities in the broader network also support research in this area. The continued growth of a large, international user community is a testament to this success and the overall utility of these facilities. Existing User Facilities are less well aligned with several of the Grand Challenges in EES science: the cryosphere (Grand Challenge 3.6), and Earth system stability and predictability (Grand Challenges 3.7 and 3.8). The EES working group developed numerous recommendations for new User Facility capacities and for new collaborations, recommendations that also address direct facility alignment with the current research portfolio and the Grand Challenges.

A number of EES recommendations (see Chapter 3. Earth and Environmental Systems, p. 22) focused on specific science aims to which User Facilities can contribute, such as atmospheric measurements using aircraft and balloon systems, aerosols and clouds, ice nucleation, and cryosphere change. Others focused on developing User Facility capacity in manipulative experiments across scales of organization, from "ecotron" chambers to field-scale experiments and from molecules and omics to Earth system science. Several recommendations involve developing observation networks of AmeriFlux omics-to-ecosystems "supersites" that could help define how processes at small scales emerge at the system scale, across the central United States to address precipitation, and to address tree structure and forest dynamics. EES recommendations also included (1) data management, computation, archival, and development; (2) cross-facility collaboration and training; and (3) additional workshops and efforts to address specific Grand Challenges. Finally, the EES working group repeated and expanded on the call articulated in the Grand Challenges report to develop a computational and synthesis User Facility focused on modeling, data-model fusion, and scaling over the full purview of BER.

Microbial to Earth System Pathways (MESP)

The Grand Challenges of MESP are to (1) define the levels of biological organization most relevant to scaling from single cells to ecosystems and global cycles; (2) capture how that organization varies in time and space; and (3) identify critical interactions that dictate rates of carbon, nutrient, and energy transformation in different environments. There is a fundamental gap at the microbial habitat scale in understanding how microbes (operating at the scale of microns to tens of microns) influence ecosystem-scale processes. Biogeochemical cycling is influenced both by microbial interactions ("bio") and the local microenvironment ("geo") that microbes experience, and user projects are contributing valuable pieces of information at this scale. However, the extraordinarily difficult tasks of integrating these data (from measurements and modeling) and developing connections across scales now need more focused attention and support.

Impediments exist to establishment and advancement of User Facility projects. Specifically, the promise of integrating process modeling with measurements and developing connections across scales will require substantial allocation of resources (i.e., personnel and equipment) for full realization. To address this need, the subcommittee recommends expanding EMSL computational support staff and their expertise to include the array of applications and codes relevant to BER users. Immediate data release rules at some User Facilities have detrimental impacts on inter-Facility and interdisciplinary projects, leading to a recommendation to institute a time delay before data are released, either until publication or for 1 year after a user project ends, whichever comes first. Also, sample throughput and data analytics have not been able to keep up with user demand for downstream data processing. One suggestion to address this gap is to shift weight toward metrics of User Facility success that recognize Facility efforts in maintaining a productive, returning user base, rather than weighting toward total numbers of users served.

To align User Facilities optimally with future needs for MESP research identified in the Grand Challenges report, specific recommendations included (1) exploring diverse scaling strategies to integrate observations and prediction, (2) developing a robust multiscale framework as a scaffold for collaboration among modelers and experimentalists, and (3) fostering interdisciplinary interactions among both established and young scientists. Recommendations for new capacities include new investments in midrange computing infrastructure *and* in personnel time to enable process modeling and data-related computation; development of a robust framework that connects and informs experimentation and modeling at multiple scales; and development of field-deployable, multimodal, remotely controlled sensors. Ideally, these sensors would conduct nondestructive measurements to characterize how microbial habitat-scale heterogeneity and dynamics influence biogeochemical processes, as well as validate the relevance of lab experiments in the field. Finally, a number of opportunities for User Facility collaboration are recommended. Collaboration among User Facilities, groups within DOE, and interagency partners such as the National Ecological Observatory Network (NEON) will be essential for tackling the very complex task of understanding links from microbial to Earth system processes. Strategies to foster collaboration could include (1) making clear that establishing connections to outside data streams is within the purview of user proposals and (2) coordinating groups at appropriate facilities to spearhead targeted research. User Facilities can also offer interdisciplinary, interactive training that supports both the development of a multiscale, collaborative framework now and the investment in critical personnel development for the future. Synthesis workshops and campaigns, short courses, and postdoctoral fellowships all could contribute. (See Chapter 4. Microbial to Earth System Pathways, p. 44, for a complete list of recommendations in this topical area.)

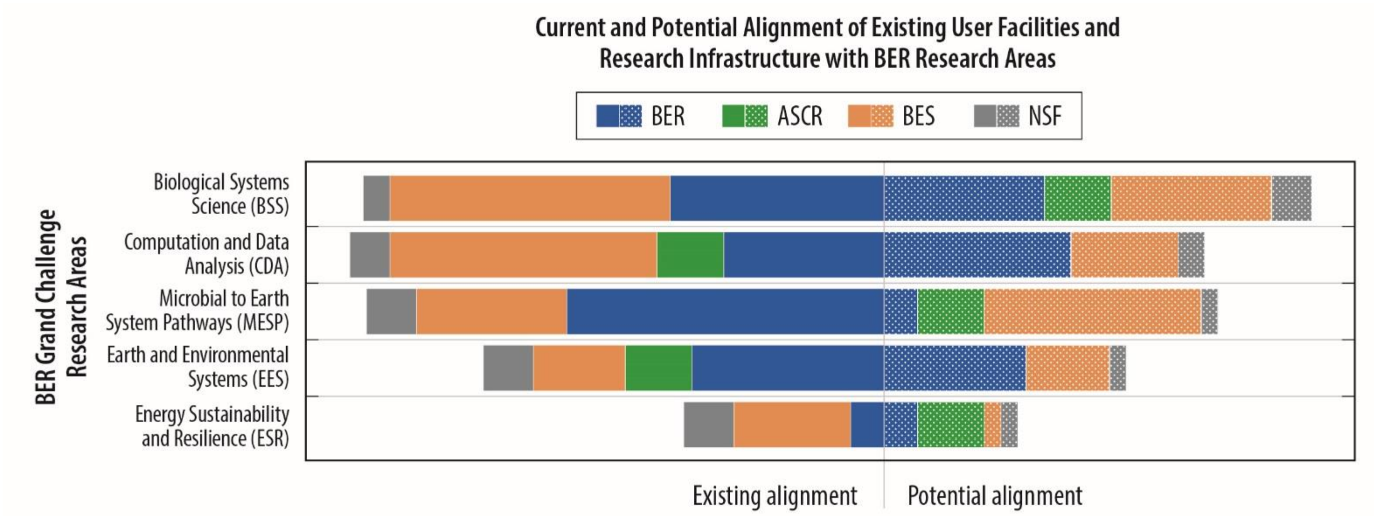

Fig. 1. User Facility Capability Alignment with BER Grand Challenges. This tornado plot summarizes the self-reported responses from User Facilities to a survey asking how strongly their current capabilities align with the Grand Challenges described in the 2017 report,[2] which are partially listed in Box 1, p. 5. The darker, more intense colored bars on the left correspond to "existing alignment" of capabilities in each of the five BER Grand Challenge research areas. Lighter colored bars on the right correspond to "potential alignment." This survey complements a similar User Facility evaluation included in Table ES-1 of the 2017 Grand Challenges report. **Key: BER**, DOE Office of Biological and Environmental Research (includes Atmospheric Radiation Measurement User Facility, Environmental Molecular Sciences Laboratory, Joint Genome Institute, AmeriFlux Network, Next-Generation Ecosystem Experiments, Spruce and Peatland Responses Under Changing Environments, and Systems Biology Knowledgebase); **ASCR**, DOE Office of Advanced Scientific Computing Research (includes computing facilities); **BES**, DOE Office of Basic Energy Sciences (includes light and neutron sources and Nanoscale Science Research Centers); **NSF**, National Science Foundation (includes National Ecological Observatory Network).

[2] BERAC. 2017. *Grand Challenges for Biological and Environmental Research: Progress and Future Vision; A Report from the Biological and Environmental Research Advisory Committee*, DOE/SC-0190, BERAC Subcommittee on Grand Research Challenges for Biological and Environmental Research (science.energy.gov/~/media/ber/berac/pdf/Reports/BERAC-2017-Grand-Challenges-Report.pdf).

Box 1. 2017 Grand Challenges (partial listing)

Biological Systems Science

2.1 Understand the biological complexity of plant and microbial metabolism and interfaces across scales spanning molecules to ecosystems.

2.2 Develop technologies to identify DOE mission–relevant metabolic capabilities and engineering possibilities in bacteria, fungi, archaea, viruses, plants, and mixed communities.

2.3 Optimize the use of large datasets that integrate omics surveys with biochemical and biophysical measurements to generate knowledge and identify biological paradigms.

2.4 Understand the links between genotype and phenotype in single but very diverse organisms and in communities of organisms that interact in terrestrial ecosystems.

2.5 Effectively exploit new and emerging technologies in systems biology and physical measurements (e.g., miniaturization) to accelerate biological discoveries.

Earth and Environmental Systems

3.1 Advance Earth system modeling using a hierarchy of models, from process-resolving coupled models to reduced-order models, to transform understanding of the coupled Earth system and to produce useful and credible simulations and predictions of Earth system behavior at multiple timescales.

3.2 Establish new observational technologies and use them to understand human and Earth system processes, such as land-atmosphere interactions, biogeochemical cycles, and subsurface soils, to estimate critical process parameters using novel analysis methods, such as machine learning and data science, and to quantify model errors.

3.3 Advance basic knowledge and scale-aware simulation capability for Earth system feedbacks associated with aerosols and moist processes to better quantify aerosol forcing, precipitation changes, and extreme events with consequences for energy and water cycles, global distribution of nutrients, and human health.

3.4 Advance modeling and understanding of important ecological, biological, and carbon cycle interactions and feedbacks in the climate system to identify potential tipping points and possible energy strategies.

3.5 Characterize, understand, and model the complex, multiscale water cycle processes in the Earth system including the subsurface to understand and predict water availability and human system response to extremes.

3.6 Understand the time-dependent processes and mechanisms associated with melting glaciers, ice caps, and ice sheets and their contributions to regional sea level rise.

3.7 Quantify the interplay between internally generated climate variability and externally forced response involving anthropogenic and natural factors and their relative roles in the time evolution of regional variability to understand predictability of the Earth system.

3.8 Understand the long-term Earth system stability in response to possible future Earth system outcomes and address the level of confidence and identify emergent constraints for the range of model projections.

Microbial to Earth System Pathways

4.1 Characterize the biogeochemical exchanges driven by food web and plant-microbe interactions and evaluate their process-level impacts, sensitivity to disturbances, and shifting resource availability under changing environmental regimes.

4.2 Define the sphere of influence and key elements of microbial communities in space and time relevant for predicting larger-scale ecosystem phenomena for Earth system understanding.

4.3 Integrate molecular and process data to improve the ability to define ecologically significant traits of individual taxa and communities and use trait-based models to develop predictive links between community dynamics and ecosystem processes.

4.4 Align and deepen connections among conceptual understanding, measurements, and models related to the roles of microbes in determining the rate of transformation, uptake, and loss of chemical elements from ecosystems.

Energy Sustainability and Resilience

5.1 Further develop the science of coupling energy, water, and land use across different spatial and temporal scales to understand environmental impacts and changing climate and to better predict net biogeochemical fluxes.

5.2 Use observational, experimental, and model-based approaches to explore the sustainability of alternative energy systems, incorporating stability and resilience analysis, uncertainty, transition paths from current infrastructures, and the use of appropriate common metrics.

5.3 Understand how variability and change in natural systems affect energy system structure and function and determine how best to build this knowledge into models.

5.4 Create new data streams and more effectively use existing observations to ensure the availability of scale-appropriate data, particularly related to high-resolution land use, landscape infrastructure, demographic change, and energy-land-water research.

Computation and Data Analysis

6.1 Develop robust approaches for large-scale data collection, curation, annotation, and maintenance.

6.2 Develop computing and software infrastructure to enable large-scale data (i.e., Big Data) storage and analysis.

6.3 Conduct research to develop suitable algorithms and programming models that can harness current and future computer architectures to effectively model complex coupled systems and analyze extreme-scale data.

6.4 Engineer advanced computational modeling combined with data integration across temporal and spatial scales.

6.5 Conduct research and develop activities that support human understanding of large-scale, multimodal data streams, including the ability to steer experiments in real time.

Energy Sustainability and Resilience (ESR)

Grand Challenges in ESR are large and, with some exceptions, not within the purview of research areas traditionally addressed by BER User Facilities. Nor, as the survey showed, are they clearly supported by the network of User Facilities in the broader community. Figure 1, p. 4, highlights a gap between the existing capabilities at User Facilities and the ESR Grand Challenges. For this research area, alignment is weak for existing BER User Facilities as well as for User Facilities in the broader network surveyed (beyond BER). This gap reiterates findings from the Grand Challenges report as well—Table ES-1 in that document also indicates weak alignment between Grand Challenges in energy sustainability and existing BER User Facilities. For this reason, the ESR team focused on a visionary idea for a network of research centers and on a single recommendation to establish a strategically distributed network of research centers focusing on the Science of Energy and Environmental Resilience (SEER), including a central Coordination, Integration, and Visualization (CIV) Center and User Facility, which would be responsible for coordinating, synthesizing, and increasing the impact of the distributed SEER Centers. This idea builds on multiple past assessments and recommendations. The SEER network would fill a vital national need by dramatically improving the scientific understanding of how the nation's co-evolving human and natural systems, especially those related to energy production and use, are changing across different geographic contexts and spatial and temporal scales. Collectively, the SEER and CIV centers would dramatically enhance the ability of DOE to meet its mission to ensure the country's security and prosperity by addressing U.S. energy and environmental challenges through transformative science and technology solutions. (See Chapter 5. Energy Sustainability and Resilience, p. 56, for a complete list of recommendations in this topical area.)

Computation and Data Analysis (CDA)

CDA resources are vital in the biological and environmental sciences, and DOE's User Facilities house tremendous assets in this area. The survey of User Facility leadership and user groups indicates strong and potential alignment with the Grand Challenges in all areas (see Figure 1). The CDA working group identified key areas for improvement—supporting the existing research portfolio, addressing the Grand Challenges, developing new capacity, and improving collaborations (see Chapter 6. Computation and Data Analysis, p. 67). Other key areas identified include the needs for storage and management of raw and derived data; real-time streaming and interactive computing; support for complex workflows; long-term software maintenance; access to testbeds and training; and metadata management and standardization, a call echoed in other chapters of this document. In new capacity development and collaboration, Chapter 6 articulates the value of a federated organization of computing resources (i.e., centralized, but with autonomy), the need for new resources in midrange computing, and the need to secure data preservation. Some recommendations focused specifically on the need for BER to develop a strategic approach to computing needs, such as in infrastructure, applications, usage policies, and intra-agency coordination. Reflecting the central and essential role of computation and data analysis for BER research, many additional recommendations for collaboration were identified—for mechanisms to facilitate research interactions between data scientists (informaticists, analysts, and statisticians) and biologists and environmental scientists, for domain-specific coordination across BER, and for various collaborations across the DOE Office of Science. (See also Appendix C. BERAC Subcommittee Workshop Agenda, p. 92; Appendix D. BERAC Members and Workshop Participants, p. 93; and Appendix E. Acronyms and Abbreviations, p. 96.)

Chapter 2. Biological Systems Science

In seeking to understand the genome-encoded properties of plants and microbes, BER seeks to further develop their potential for redesign for beneficial purposes. The current emphasis is on understanding microbes and plants with characteristics suitable for the production of fuels and chemical products from renewable biomass. BER's Biological Systems Science Division (BSSD) supports research directed at understanding the complex processes and structures underlying the organisms to be engineered. The challenges of understanding these biological systems, their metabolic pathways, and their interdependencies were recognized in the 2017 Grand Challenges report[3]: "Greater insights are needed into the regulation of these pathways, the genes responsible for the reactions, and environmental influences on the reactions. This improved understanding is a precursor to enabling changes in pathways that may uncover new or more efficient energy sources." Current barriers to achieving this understanding include fragmented access to available methods and cumbersome protocols.

The authors of the 2017 Grand Challenges document recognized the importance of the BER-supported User Facilities and their counterparts supported through other parts of DOE. The organic growth of such interactions is under way, as exemplified by the Facilities Integrating Collaborations for User Science (FICUS) initiative between the Joint Genome Institute (JGI) and Environmental Molecular Sciences Laboratory (EMSL). This collaboration is one of the best established, and discussions are under way to expand it to existing synchrotron structural biology infrastructures. Other ongoing examples include the use of small-angle neutron and X-ray scattering (SANS and SAXS) resources between Oak Ridge National Laboratory (ORNL) and Brookhaven National Laboratory (BNL) and the growth of EMSL's interaction with light and neutron sources. BERAC supports these activities and encourages stronger and deeper interaction as the realization of the multiple benefits from integrated approaches allows investigation of complex biology.

CHARGE 1 RESPONSE

Alignment of User Facilities to Current Biological Systems Science Research

The BER-supported User Facilities, and User Facilities throughout the country that contribute to Biological Systems Science (BSS)–related goals, are already advancing science at the cutting edge of the BER BSS portfolio. A detailed understanding of metabolism requires imaging; comprehensive measurements of the metabolome, fluxome, proteome, and transcriptome; and modeling. Existing BER User Facilities and modes of research excel in these areas. Furthermore, state-of-the-art technologies currently available at the User Facilities can sequence and annotate organism genomes (JGI); perturb metabolic pathways using synthetic biology (JGI); and study the temporal and spatial organization of metabolism, including the subcellular localization of metabolic pathways, the manner in which the activity of a pathway varies with time and regulation of metabolism at the transcriptomic level (EMSL), and structure-to-function relationships in metabolism (EMSL and BNL). Resources such as DOE's Systems Biology Knowledgebase (KBase) provide a platform that can be used to develop predictive models of metabolism by integration of data from various sources, helping to fill knowledge gaps and allow the generation of hypotheses that

[3] BERAC. 2017. *Grand Challenges for Biological and Environmental Research: Progress and Future Vision; A Report from the Biological and Environmental Research Advisory Committee*, DOE/SC-0190, BERAC Subcommittee on Grand Research Challenges for Biological and Environmental Research (science.energy.gov/~/media/ber/berac/pdf/Reports/ BERAC-2017-Grand-Challenges-Report.pdf).

inform future rounds of experiments. The following are examples of current User Facility technologies and capabilities that are advancing or are well positioned to advance key BSS research areas.

Discovering Metabolic Pathways

Understanding metabolic pathways and their roles in cellular, organismal, and community metabolism is a fundamental goal in BSS research, and it relies on a diverse collection of organisms in which metabolic pathways can be explored, manipulated, and engineered. Developing universal tools to advance this science is critical, and recent work by JGI to develop a universal strain-engineering platform (Chassis-Independent Recombination Assisted Genome Engineering (CRAGE) shows promise as such a system. JGI researchers have used this system to engineer >40 species of bacteria, as well as to deliver large amounts of DNA of up to approximately 60 kilobases (kb), which can contain the genes encoding enzymes that define entire biosynthetic pathways. Further development of this technology, and a pipeline to automate the process, could provide the foundation for a wide range of new chassis for metabolism discovery and optimization.

Integrated Omics

BER User Facilities are at the forefront of integrating genome-wide datasets to better predict the function of previously unannotated or poorly annotated genes. Transcriptomic datasets are increasingly common, while proteomic and metabolomic datasets are less so because of greater technical challenges in generating them. Subjecting two or all three of these types of data to the same perturbations is highly unusual, but this multiomic approach has the potential to go beyond the predictions of transcriptional datasets to show changes in cellular activities, not just information coding. Furthermore, having such datasets with high temporal and spatial resolution is exceedingly rare. DOE's BER and Basic Energy Sciences (BES) facilities have expertise in generating all three of these types of datasets. For example, EMSL's capability to co-localize transcriptomic, proteomic, and metabolomic changes allows for integrative multiomic sampling either from static or dynamic experiments. Such facilities need to be maintained and expanded to help generate additional parallel datasets. Also critical are the additional human resources needed to develop approaches to integrate these datasets and the communication networks to combine efforts and datasets among laboratories. Meeting these needs will require overarching strategies, such as a relatively simple approach for common naming of samples for comparison between labs and the more complex development of software and algorithms to overlay these datasets.

A critical remaining challenge is to express sufficient protein for functional annotation. The ability to synthesize complete genes has helped increase the rate of these experiments, yet these approaches still require the development of highly efficient and effective expression platforms to synthesize sufficient quantities of encoded proteins to define their structure and function. Levels of protein synthesized from these systems are often insufficient for defining protein structure. Similarly, high-throughput cell-free systems that can be used to rapidly prototype metabolic pathways to identify metabolite products and intermediates are also limited by the yields of protein. EMSL has recently developed a cell-free pipeline to improve structural analysis of interesting protein targets identified from time-resolved or dynamic proteomic studies. In parallel, JGI has recently invested in a new Emerging Technologies Opportunity Program (ETOP) for developing cell-free systems optimized for particular metabolic nodes. If successful, this project will deliver and demonstrate a valuable platform for high-throughput large-scale metabolic exploration.

Genotype to Phenotype

A central challenge for all biologists is understanding how the genotype of organisms defines their phenotype. This question is also central to the BER goals of understanding biological systems at levels ranging from how specific genes define cellular activities, cellular phenotypes, and organisms' phenotypes to how environmental changes and community structure of terrestrial ecosystems influence phenotypes of a diverse array of organisms that make up these communities. The DOE User Facilities are currently well equipped to examine these relationships for individual genes and how they affect a diversity of individual phenotypes including transcriptional, metabolic, and structural phenotypes. The JGI houses state-of-the-art capabilities in high-throughput sequencing, functional genomics, DNA synthesis, and metabolomics, all underpinned by high-performance computing (HPC). Ongoing work at EMSL is poised to integrate genomics with dynamic, live-cell, environmentally controlled, and time-resolved studies using a suite of modalities. JGI and its partners at Lawrence Berkeley National Laboratory (LBNL) recently published a comprehensive pipeline designed for functional annotation in microorganisms, the *Functional Encyclopedia of Bacteria and Archaea* (FEBA), which is a system for discovering the function of protein products from genes that are not readily predicted from their sequence. The plan is to embed this pipeline into KBase as a powerful tool in the quest to infer gene function and metabolic networks and, in general, to link genotype to phenotype. This accomplishment sets the stage for more strategic advances in this arena. This capacity in the User Facilities is very strong and warrants continued support and development.

CHARGE 2 RESPONSE

Alignment of User Facilities to Address Future Needs and Grand Challenges in Biological Systems Science

The Grand Challenge document outlines five Grand challenges for systems biology (see BERAC 2017, p. 4). As a whole, these Grand Challenges (included herein) recognize the interconnections among underlying biological complexity, the technology developments necessary to generate new knowledge, and the scale of these datasets. The next technical steps in achieving Grand Challenges 2.1 through 2.5 can be divided into the following five experimental BSS objectives whose achievement will depend on User Facilities.

1. Establish new paradigms to transition from unannotated gene sequence to experimentally validated functional annotation.
2. Develop a deeper knowledge of the spatial and temporal control of metabolism.
3. Create the technology to enable rapid whole organism phenotyping.
4. Explore the interplay between metabolism, signaling, and the response of organisms and communities to their environment.
5. Harmonize the data produced across User Facilities and scales.

The sections that follow describe BSS research needs and knowledge gaps for each of the five Grand Challenges and their related objectives and provide BERAC's specific recommendations aimed toward enabling BER to realize these ambitious goals.

Grand Challenge 2.1: Understand the biological complexity of plant and microbial metabolism and interfaces across scales spanning molecules to ecosystems.

Developing Databases, Libraries, and Models for Metabolism

Several major hurdles need to be overcome to strengthen the understanding of metabolism. Of prime importance are experimentally based annotation of gene function (Objective 1) and development of curated metabolic pathway databases and stoichiometric metabolic models (or "flux-balance-analysis-ready" models) corresponding to the annotation (Objective 5). A difficulty encountered in doing this is the determination of the intracellular organization of metabolites, enzymes, and pathways in eukaryotes, especially plants. In this context, the integration of data from transit peptide sequences, imaging, metabolomics, and fluxomics will serve to develop maps of subcellular organization of metabolism (Objective 2). Another significant goal is the establishment and availability of comprehensive structural libraries of metabolites and enzymes (Objectives 2, 3, 4), which are currently lacking. Concurrent with this, methodologies that identify metabolites on the basis of structure, metabolic pathways, and information present in existing libraries are highly desirable.

Measuring Cellular-Level Metabolism

Following the establishment of metabolic databases, models with subcellular organization, and metabolite libraries, the next natural step is the directed measurement of metabolic fluxes, between organelles and between species interacting in a community. This can be accomplished by combining imaging, isotope labeling, and modeling (Objectives 2, 3, 4). The complex nature of isotopomer networks has currently limited isotope-based flux measurements to relatively homogeneous and steady-state systems. User Facilities can drive progress by developing computational methods that expand these models to relax the simplistic assumptions, as well as acquisition of technologies that support measuring fluxes under such conditions, *in situ*. These methods can also be applied to environmental samples, which will also be of interest to BER. Furthermore, molecular dynamics simulations, which seek to develop a molecular understanding and predictions of membrane transport (Objectives 2, 5), need to be integrated with these measurements. Supporting User Facilities equipped for the measurement of metabolic and signaling interactions between cells and within communities in a dynamic manner is a worthy long-term goal (Objectives 2, 3, 4). This will entail further development of capabilities for *in situ* measurements (Objectives 2, 3, 4) and single-cell measurements (Objectives 2, 3, 4) in addition to the goals listed above.

Moving Beyond the Rate-Limiting Steps of Metabolic Modeling

The rate-limiting step in understanding metabolism is the development of comprehensive, predictive models, applicable to model systems and to natural ecosystems (Objectives 2, 3, 4, 5). While stoichiometric metabolic models for performing flux balance analysis are available or being developed for many species, their analysis requires understanding why and how pathways are controlled. Additionally, stoichiometric models can be combined with omic data (transcriptomic and fluxomic) to predict the mechanisms by which changes in metabolites occur. Methods for this integration of omics and for linking them to modeling are now emerging. Kinetic modeling of metabolism and its integration into the aforementioned measurement and modeling frameworks needs much development.

Scaling from Model Systems to Ecosystems

Research in biological systems science supported by BER and other agencies has been tremendously successful in understanding organismal biology, biochemistry, and its genetic foundation, focusing on organisms in isolation or in simple constructed communities. A new frontier for biological systems science captured in Grand Challenge 2.1 recognizes a key frontier in understanding metabolism across scales, from molecules to ecosystems. Tools are developing that are able to quantify metabolism and element assimilation by microorganisms in complex communities, such as metabolic flux modeling and coupling omics with stable isotopes. Both meet this goal to extend biological systems science from molecules to ecosystems, and they are areas where investments by User Facilities are likely to be strategic.

We recommend that JGI and EMSL evaluate establishing a user capacity in stable isotope probing of environmental samples, one that includes standards for isotope enrichment, pipelines for quantitative analysis, and engineering to improve precision. This effort could build effectively on existing projects in JGI's ETOP portfolio of projects, where, for example, techniques for more standardized and precise isopycnic separation of nucleic acids are being developed. We also recommend that the BER KBase community explore developing KBase as a platform for the quantitative interpretation of isotopomer-enabled metabolic flux modeling of central metabolism, and specific metabolic pathways of biogeochemical interest.

Training an Interdisciplinary Workforce to Understand Metabolism

Finally, the interdisciplinary nature of metabolism research makes it necessary to focus on workforce development for multiscale thinking (Objective 5). This need requires training of biologists, analytical chemists, engineers, physical scientists, computer scientists, mathematicians, and statisticians to collaborate toward 21st century models of metabolism. Conferences as well as short- and long-term training programs and projects that span disciplines will contribute extensively to dialogue across these disciplines.

Recommendations

2.1 Develop metabolic pathway databases based on experimentally annotated gene function and integrate metabolic data needed to achieve subcellular organization of metabolites, enzymes, and pathways.

2.2 Develop structural libraries for metabolites and enzymes.

2.3 Obtain equipment designed to dynamically measure intracellular and interspecies fluxes of metabolism and transport by developing imaging and isotope labeling technologies, applicable to organisms interacting in complex communities, *in situ.*

2.4 Develop methods for *in situ* measurements and single-cell measurements.

2.5 Integrate molecular dynamics simulations into flux measurements.

2.6 Develop stoichiometric and kinetic models of metabolism that integrate omic data and allow the transition from observations of changes in gene expression to metabolic activity.

2.7 Establish capacity at JGI or EMSL for stable isotope probing.

2.8 Develop KBase to support quantitative interpretation of isotopomer-enabled metabolic flux modeling of central metabolism, and other specific metabolic pathways.

2.9 Train an interdisciplinary workforce for improving the understanding of metabolism.

Grand Challenge 2.2: Develop technologies to identify DOE mission–relevant metabolic capabilities and engineering possibilities in bacteria, fungi, archaea, viruses, plants, and mixed communities.

Deploying Synthetic Biology to Efficiently Produce Products

To gain understanding of metabolic pathways and their interplay in cellular, organismal, and community metabolism, technical approaches to perturb and measure metabolism are required, as indicated in the 2017 Grand Challenges report. Today, DNA sequencing significantly outpaces the ability to assign experimentally validated function to genes. The application of synthetic biology, utilizing principles from engineering and nanotechnology, affords the rational design, engineering, and iterative testing and learning necessary to systematically explore metabolism and assign function. Current challenges with synthetic biology are the limited set of available host organisms, difficulties in large DNA construct design and assembly, lack of rapid prototyping systems, and integrated measurement systems. Through expansion of the DNA Synthesis Science Program at JGI and partnerships with other User Facilities under this coordinated network, these limitations can be addressed.

Transforming Recalcitrant Strains

One difficulty in experimentally characterizing metabolism and metabolic pathways is the limited availability of genetically tractable host chassis organisms. Today, only a narrow set of organisms has been developed and explored for expression or deletion of metabolic pathways due to the lack of universal tools for strain manipulation and challenges in successfully transforming organisms due to restriction modification systems. Therefore, development and deployment of technologies to engineer previously genetically intractable organisms to expand the repertoire of genetic chassis, including bacteria, archaea, viruses, fungi, and plants, are required (Objectives 1, 3, 4). Building on systems like CRAGE, developed at JGI, the User Facilities can help develop new methods for facile porting of biosynthetic pathways between organisms to investigate the role of physiological context on gene expression and metabolism (Objectives 1, 3, 4).

Applying Gene Editing to Diverse Microbial and Plant Species

Traditionally, most experimental investigations focus on a limited set of genes or metabolic pathways that provide a narrow understanding of integration of entire pathways or processes. To explore metabolism and assign function, large-scale genome-wide approaches are required to manipulate expression of genes. The recent advent of CRISPR gene-editing technology offers one such method to up- and down-regulate genes using libraries of guide RNAs and single- and multiplex approaches. To successfully achieve this genome-wide gene editing for a large diversity of microbial and plant species, highly efficient delivery systems for CRISPR and guide RNAs are desired that are based on either separate nuclease/guide RNA expression or delivery of nucleoprotein complexes. Further development of such systems is required. Alternative genome-wide gene disruption technology, such as Tn-Seq, enables creation of genome-wide transposition barcode-labeled libraries that can be screened under a variety of conditions. JGI, and its partners at LBNL, recently published a comprehensive study in *Nature* about FEBA. This type of technique should be extended to other organisms (Objectives 1, 2, 3, 4).

Rapidly Synthesizing *De Novo* DNA and Assembling Large DNA Molecules

The central schema for synthetic biology is the design-build-test-learn (DBTL) cycle. The efficiency of this schema depends on highly efficient workflow schemes and automation that enable the individual steps. Various iterations of platforms exist in both academia and industry, but they have not been standardized. To accelerate the DBTL cycle, advancements in optimized design, facile DNA assembly, high-throughput assay platforms, and high-performance computational analysis, simulation, and modeling are needed. The JGI, with its partners, is well poised to develop such accelerated and optimized platforms.

DNA synthesis is largely based on older chemical methods that are only now being scaled to a point where costs per base pair are <$0.10. This cost is still prohibitive for many large-scale experiments that require large volumes of synthesized DNA. New methods are required for DNA synthesis that further reduce cost. Improvement of high-throughput methods for design, build, and assembly of large DNA constructs that encompass complete biosynthetic pathways is also required (Objectives 1, 3, 4).

As noted above (see Integrated Omics section, p. 8), there is strong existing capacity for developing a platform for high-throughput and large-scale metabolic exploration via cell-free pipelines: EMSL's cell-free pipeline for structural analysis of protein targets, and JGI's ETOP for developing cell-free systems optimized for metabolism. Continued support of these programs will help address Grand Challenge 2.2.

Further opportunities in optimization of DBTL exist in the test and learn parts in which functional understanding can be gained. Integrated multiomics technologies should be deployed to measure regulation of metabolism and effects of perturbing metabolism using engineered organisms (Objectives 1, 2, 3, 4). Cellular sensors should be developed to report on the metabolic state *in situ* without altering the cell or reaction flow. Regulatory networks should be developed that ensure optimal expression of metabolic pathways under target conditions. Organisms with engineered metabolic pathways can be combined in model ecosystems to study effects on community metabolism and organism interaction (Objectives 3, 4). Data from engineered metabolism studies should be incorporated into metabolic models that can ultimately predict metabolic pathways and perturbation effects (Objective 5).

Recommendations

2.10 Improve methods for large DNA construct design and assembly and expand the availability of rapid prototyping systems.

2.11 Develop and deploy technologies to engineer previously genetically intractable organisms.

2.12 Develop new methods for facile porting of biosynthetic pathways between organisms to investigate the role of physiological context in gene expression and metabolism.

2.13 Expand the tools available for genome-wide genetic disruption and their application to a range of organisms.

2.14 Develop cellular sensors for monitoring metabolism and metabolic state in organisms and how they are influenced by their ecosystem.

2.15 Enhance capability for *de novo* DNA synthesis and assembly of large DNA molecules.

Grand Challenge 2.3: Optimize the use of large datasets that integrate omics surveys with biochemical and biophysical measurements to generate knowledge and identify biological paradigms.

Organizing, Archiving, and Retrieving Data

Readily apparent from the 2017 Grand Challenges report is that more data must be generated to achieve the stated goals. Therefore, facilities require the infrastructure to store and retrieve the datasets for subsequent use. Each User Facility needs a space for data storage where information can be archived and retrieved on demand. DOE'S Office of Advanced Scientific Computing Research (ASCR) facilities have enormous archive resources in the form of high-performance tape systems that can be accessed directly via Globus for high-speed transfer via ESnet. The growth in data generation at facilities such as the light sources indicates the need for larger tape and archive resources and is part of the planning for ASCR facilities. One model that has worked well at DOE's National Energy Research Scientific Computing (NERSC) Center is called "sponsored storage," where users or facilities can procure tape that can be housed in perpetuity.

Measuring Data Impact for Cost Efficiency

Each facility also needs a way to monitor the access and retrieval of its datasets in order to quantify the number of downloads, number of users accessing a particular data resource, and whether or not these data continue to be of value over time. This translates to a hierarchical data management strategy, where frequently accessed information is kept in cache and less frequently accessed data are in cold storage. Scientists must be challenged to think of ways to maintain only the data necessary to reproduce a result.

Data Search and Access Capabilities

The ability to search large datasets is a challenging problem that requires rich contextual information about the data that the user is trying to retrieve. As the number and variety of datasets increase, it becomes harder to specify a unique set of terms for finding the dataset of interest. There have been a number of efforts to consolidate data search efforts in industry and academia. An interesting example is the Repositive resource, which has created unified access to thousands of human genomic datasets. DOE needs to build similar infrastructure for environmental omics datasets, and, as a start, DOE's JGI, EMSL, and KBase should create a shared search platform.

Coordinating Sequence Datasets with Their Associated Chemical, Physical, Temporal, and Environmental Treatment Datasets

This challenge is similar in spirit to the search and retrieval problem. The rich context or the chemical, physical, temporal, and environmental treatments can provide additional constraints that enable scientists to retrieve the most relevant datasets for their question.

The User Facilities can assist in this effort by requiring minimum sets of metadata for any samples that are sent to the facilities for processing and by having equivalent requests across all facilities. For example, experimental apparatus such as an EcoFAB can provide controlled environments for collecting experimental data and also determining the minimum sets of metadata that are needed to allow broad reproducibility and reliable correlation of multimodal data. These proofs of concept can be used to demonstrate the utility of metadata collection to answer questions. This approach needs to be taken further to allow for similarity-based searches since it is unreasonable to think that all metadata will have a perfect match across samples, analysis systems, and organizations.

These data should be immediately added to a system that can be ingested by the search platform proposed above. It is the responsibility of the researcher to ensure that the metadata are provided. Integration tests must be run to ensure that the new data can be associated with other experiments, and the researcher should validate that the associations are correct.

Providing Quality Control and Functional Annotation

Robust search and exploration of the available data will open up the possibility of crowd-sourced quality control and annotation of the existing annotated data. The current state-of-the-art methodology involves the comparison of sequence data to databases of "known organisms." These databases are known to contain errors, and these errors propagate when sequences are linked back to sequences with errors. A platform that enables community-driven consensus is needed so that scientists from across domains and disciplines can view and assess data generated by the User Facilities. These data users can contribute their expertise outside the traditional facility engagement mechanisms.

Recommendations

2.16 Build a highly dynamic, shared search platform among JGI, EMSL, and KBase to enable data correction and method sharing outside of static publications.

2.17 Establish and adhere to metadata standards for JGI, EMSL, and KBase and lead efforts in setting such standards in collaboration with other large scientific organizations.

2.18 Leverage ASCR compute resources for data storage and large-scale computing.

2.19 Issue joint funding calls with ASCR to encourage collaboration among biologists, mathematicians, and computer scientists on the development of methods for multimodal data integration and understandable machine learning.

Grand Challenge 2.4: Understand the links between genotype and phenotype in single but very diverse organisms and in communities of organisms that interact in terrestrial ecosystems.

Understanding how the genotype defines the phenotype is central to systems biology. The gene-by-gene approach currently practiced at JGI has achieved important results and insights. The next step is to expand this effort to be more rapid, providing the genome-wide understanding that is needed. Integrating transcriptomics, proteomics, metabolomics, and high-throughput approaches for imaging and growth and developmental insight can provide understanding of the genome-wide controls of phenotype. Another important need is to go beyond understanding the linkage between genotype and phenotypes under one environmental condition in one cell type, to understand how cellular activities and whole-organism phenotypes change across space and time both within individual organisms and across whole communities. Ongoing work at EMSL already provides access to researchers for pursuing these dynamic, live-cell, environmentally controlled and time-resolved studies using a suite of modalities. However, improving throughput and multimodal integration of these approaches will further expand the benefit to users.

Specifically, DOE User Facilities should (1) acquire imaging and mass spectroscopy equipment, as well as facilities for controlled growth of organisms for these analyses (incubator, growth chambers, and plant greenhouses) to allow high-throughput analysis and visualization of transcriptional, metabolic, and developmental phenotypes; (2) hire researchers and programmers to develop approaches to more seamlessly integrate multimodal and multiscale results; and (3) establish additional facilities or capabilities that allow controlled perturbation of growth conditions of individual organisms and communities as outlined below.

Developing a Hierarchical Annotation Pipeline Integrating Experimental and Computational Approaches to Assess Functional Annotation Quality

Transcriptional datasets are now commonplace, but our ability to interpret them is limited by incomplete annotations. Even in the best characterized species, such as in *Arabidopsis*, ~25% of the genes are not annotated and many of the others are predicted, without functional tests.[4] Wild and crop species are far behind this number. *Computational methods to better predict function are needed, as are high-throughput methods to validate function*. User Facilities with dedicated bioinformatic staff are in a unique

[4] Bolger, M. E., B. Arsova, and B. Usadel. 2018. "Plant Genome and Transcriptome Annotations: From Misconceptions to Simple Solutions," *Briefings in Bioinformatics* **19**(3), 437–49. DOI:10.1093/bib/bbw135.

position to develop algorithms to predict gene function and to combine these sequence-based approaches with experimental tests of function. *We recommend that additional human resources are needed to support these time-intensive computational efforts*. High-throughput analyses that test the function of these predictions are also required to solidify the functional relationships. Some of these approaches are in place in the users' labs, but strengthening and expanding these efforts are needed as outlined below.

Applying Multiple Functional Genomics Approaches to Single Samples and Cells

Two approaches are most commonly employed in functional genomics. The first is a biochemical approach in which genes can be cloned into expression vectors and proteins produced and function tested in biochemical assays. This works well for proteins with predicted function that are straightforward to analyze, such as enzymatic activity or high-throughput binding assays, and *pipelines for testing protein function or determining protein structure are in place at some User Facilities*. The current User Facilities, which are limited in the number of genes that can be tested and in the diversity of functional assays, should be expanded. Similarly, BER scientists can already take advantage of synchrotron-based structural biology beamlines or nuclear magnetic resonance (NMR) capabilities at EMSL, but access to high-resolution cryo-electron microscopy (cryo-EM), which has less-stringent sample requirements, remains limited. Expanding the availability of high-throughput cryo-EM capabilities available to BER users by increasing the number of available instruments should be a focus for EMSL and DOE to accelerate structural annotation of genomic dark space.

A second powerful approach is to test protein function using insertion mutants. This approach is particularly appropriate when protein function is not predicted, or the prediction of function is tenuous. This has proven very powerful in model species, but such libraries of mutants are not available in many nonmodels. Development of such libraries would be a very powerful resource for DOE researchers and the User Facilities could work to generate such libraries for a select number of species.

Integrating Genome-Wide Datasets to Better Predict Genes with Important Functions

As described in the section titled "Alignment of User Facilities to Current Biological Systems Science Research," p. 7. BER User Facilities are at the forefront of integrating omics datasets. Nevertheless, there is substantial room for growth in this area, facilitating combined efforts between laboratories, developing software and algorithms to overlay datasets, and cultivating new personnel capable of the integration.

Expanding Facilities to Better Characterize How Changes in Genotype Lead to Changes in Phenotype of Both Communities and Individual Organisms

One of the greatest technical difficulties in addressing this Grand Challenge—understanding the relationship between genotype and phenotype—is being able to carefully and reproducibly characterize the phenotype. Accomplishing this can include imaging of signaling, measurement of growth and development, or monitoring changes in populations. Additional facilities to provide this information, especially under carefully controlled conditions and with an automated approach are needed.

Recommendations

2.20 Develop high-throughput computational methods to better predict function of gene products, including expanding the User Facility computational biology team.

2.21 Develop expression platforms capable of generating sufficient protein for characterization of protein structure and function.

2.22 Employ genome-wide gene disruption and gene expression technologies such as CRISPR, Tn-Seq, and Dub-Seq, which allow for systematic assignment of critical genes under specific conditions.

2.23 Deploy integrated multiomics technologies to understand genome-wide changes in gene expression and metabolism.

2.24 Enhance the integration of "omic" and other data generated at multiple User Facilities by enhancing the coordination among these facilities.

2.25 Develop facilities to better characterize phenotypes resulting from altered gene function, including whole-organism and population growth and development, as well as high-resolution imaging and monitoring of metabolic changes.

Grand Challenge 2.5: Effectively exploit new and emerging technologies in systems biology and physical measurements (e.g., miniaturization) to accelerate biological discoveries.

Using and Coupling Nanotechnology and Microfluidics

EMSL and JGI incorporate microfluidics and nanotechnology in several areas of their current user programs. However, new methods and approaches for exploiting and coupling nanotechnology and microfluidics more extensively into BER User Facilities are needed to expand the accessibility and impact of these techniques for high-throughput, *in situ* and single-cell applications. Advancements in fabrication, microfluidics, and nanotechnology can foster quicker screening of gene constructs, biochemical assays for proteins, performance assays for cells, assessing microbial community communication and architecture, creation of synthetic communities, observations of how cells or biosystems sense and respond to environmental perturbations, and subcellular visualization of cell dynamics. In addition, the miniaturization provided by these capabilities presents opportunities for deep phenotyping and *in situ* measurements as well as the potential to truly link omics and imaging on the same exact sample to overcome questions about potential mismatches between environmental conditions, temporal sampling, or sample preparation affecting the integration of analysis across modalities.

Leveraging the Cryo-EM Revolution

New technologies for label-free imaging and structural biology could provide foundational discoveries in BER research as described in the 2017 Grand Challenges document[5] and the report titled *Technologies for*

[5] BERAC. 2017. *Grand Challenges for Biological and Environmental Research: Progress and Future Vision; A Report from the Biological and Environmental Research Advisory Committee*, DOE/SC-0190, BERAC Subcommittee on Grand Research Challenges for Biological and Environmental Research (science.energy.gov/~/media/ber/berac/pdf/Reports/ BERAC-2017-Grand-Challenges-Report.pdf).

Characterizing Molecular and Cellular Systems Relevant to Bioenergy and Environment.[6] Although present at BER User Facilities, including EMSL, cryo-EM for single-particle structure determination and cryo-electron tomography (cryo-ET) for whole-cell three-dimensional (3D) reconstruction are underrepresented relative to their need and potential impact to BER. Expanding state-of-the-art high-throughput instrumentation and methods for electron microscopy within the BER network can lead to a wealth of new information for quantifying growth and developmental phenotypes, understanding community architecture, linking gene sequence to protein structure and function, and visualizing subcellular organization and composition to identify the localization and quantity of protein complexes within the whole-cell context. The same instruments can be used to empower microelectron diffraction for structure determination from 3D protein crystals that have proven intractable to other approaches as well as filling a critical gap for *in situ* hierarchical tomography to provide sequential 3D views across scales.

To facilitate broad impact and accessibility for BER researchers, it is recommended that BER support the deployment of cryo-EM capabilities within EMSL, an established facility that can immediately support the infrastructure needed to operate and maintain the equipment while also providing direct links to multimodal interrogation of same samples and ease of access by the broad user community. Investment in new state-of-the-art electron microscopy capabilities colocalized near other electron, ion, optical, Raman, and X-ray imaging capabilities at EMSL would advance *in situ*, dynamic, correlative, or multimodal analysis across scales. Additionally, because EMSL has recently developed a cell-free pipeline to improve structural analysis of interesting protein targets identified from time-resolved or dynamic proteomic studies, the investment in new state-of-the-art cryo-EM capabilities at EMSL could significantly accelerate both the availability and productivity of this approach to BER users. This would also facilitate electron microscopy and multiomic analysis on the same sample and provide a unique opportunity for providing direct links between whole-cell 3D nanoscale bioimaging and systems biology. An alternative approach would be to expand the available cryo-EM capabilities at a BES-operated national light source, but making these instruments focus on serving BER researchers is another approach that would accelerate gene annotation because these facilities would be equipped to probe the same sample with both X-rays and electrons to provide new mechanistic insights into the structure, dynamics, and function of individual macromolecules.

Harnessing Label-Free Imaging Approaches

Label-free imaging capabilities allow for the interrogation and tracking of biological structure, morphology, chemistry, and dynamics by detecting signals inherent to the organism of interest. In combination with other label-based and nonperturbative probes and sensors such as stable isotope labeling, selective Raman tags, and engineered fluorescent proteins, a wide array of whole-cell or organism phenotyping and characterization techniques is available to the general science community. Several of these capabilities exist at EMSL and synchrotron facilities including, but not limited to, nanoscale-secondary ion mass spectrometry (Nano-SIMS), stimulated Raman scattering (SRS) microscopy, infrared scanning near-field optical microscopy (IR-SNOM), and cryogenic soft X-ray nanotomography (cryo-SXT). Efforts should be made by the User Facilities to extend the portfolio of label-free approaches and to increase access to these capabilities for BER users. Of specific interest is the development and linkage of hybrid label-free and label-based approaches for high-throughput time-resolved imaging of live cell dynamics in controlled microfluidic environments to accelerate phenotyping efforts and visualize metabolic flow.

---

[6] U.S. DOE. 2017. *Technologies for Characterizing Molecular and Cellular Systems Relevant to Bioenergy and Environment*, DOE/SC-0189, U.S. Department of Energy Office of Science (science.energy.gov/ber/community-resources/).

Recommendations

2.26 Expand the accessibility of microfluidic and nanotechnology techniques for high-throughput, *in situ*, deep phenotyping and single-cell applications.

2.27 Develop facilities so that researchers can perform imaging on a sample and then subject these samples to omics approaches.

2.28 Invest in state-of-the-art high-throughput cryo-EM instrumentation and couple with cell-free expression capabilities within the BER network to facilitate rapid structure determination and protein annotation.

2.29 Deploy new cryo-ET capabilities within already established User Facilities for multimodal interrogation of whole cells and ease of access by the broad user community.

2.30 Extend the portfolio of microscopic imaging facilities designed to perform label-free imaging available to BER users.

CHARGE 3 RESPONSE

Development of Additional User Facility Capabilities

Many of the recommendations made under other charge categories (above) involve developing additional User Facility capabilities. For example, see those recommendations addressing Grand Challenges 2.1, p. 9, and 2.5, p. 17, where specific capacities are described and recommended. Another more general area of capacity development involves the *Recruitment and Retention of Top Talent*. To successfully address these Grand Challenges through the recommended actions, highly skilled personnel are required at the User Facilities. Current staff should be properly trained in the necessary technical disciplines. Where necessary expertise is missing, strategic hiring should be conducted to fill gaps; this may be especially critical for synthetic biologists, data scientists, analysts, and software engineers due to the highly competitive employment environment for these skills. Given the highly collaborative nature of the scientific activities, it is essential that personnel are trained in effective communication, project management, and in team science.

CHARGE 4 RESPONSE

Opportunities for Collaboration Among User Facilities

Addressing Grand Challenges requires imagination, perseverance, and detailed measurements. For the Biological Systems Science Division (BSSD) of BER, the infrastructure to enable the successful investigation of the opportunities recognized in the Grand Challenges Report[7] will be enabled through closer cooperation between the User Facilities and by making these facilities more accessible to the research community. These benefits can be obtained both by better integration of BER's user programs at EMSL and JGI and by programs that support research projects and approaches between research at BER User Facilities with those supported by DOE BES (light and neutron sources) or by NERSC. Such integration will deliver greater opportunity for impactful science and will leverage the investments made at these

[7] BERAC. 2017. *Grand Challenges for Biological and Environmental Research: Progress and Future Vision; A Report from the Biological and Environmental Research Advisory Committee*, DOE/SC-0190, BERAC Subcommittee on Grand Research Challenges for Biological and Environmental Research (science.energy.gov/~/media/ber/berac/pdf/Reports/ BERAC-2017-Grand-Challenges-Report.pdf).

facilities. By establishing an interconnected set of "centers of excellence," the return on investments will be greater while avoiding excessive duplication of resources.

BSS research is fundamentally integrative. The ability to predict how an organism will respond to environmental change (e.g., its change in fitness, the production of particular molecules, and its ability to reproduce) integrates across scales of organization within the research domain of BSS, and this ability is one of the key Grand Challenges BSSD strives to address. Answering this question is a key challenge for the future development of many aspects of the BER research portfolio. Today, even a "simple" problem such as predicting the growth of a bacterial strain under different laboratory growth conditions remains a challenge, and we assert that to understand how to make such a system predictive will require multimodal, coordinated measurements at many centers. Such integration also will be critical in achieving the vision of generating an atomic-level model of cell organization and pathways to refine theory and simulations and ultimately will lead to a "virtual cell."

Growth, gene and protein expression, protein structure, molecular characterization, exo- and endo-metabolomics, and imaging of subcellular, cellular, and supercellular structure will all play a role in unravelling the processes involved. Advancing this science requires these measurements, probably from different centers, across conditions and genetic variations. While facilities currently provide some of these resources, programs to coordinate these measurements with well-defined formal experimental designs are not well established yet.

An enhanced network across existing User Facilities within and outside BER (e.g., JGI, EMSL, NERSC, KBase, neutron sources, and light sources) presents a strong opportunity to advance these integrative goals in BER-supported BSS research. The recommendation to establish this network, called the Coordinated Network for Systems Biology (CNSB), contributes directly to aligning User Facilities with the current BER research portfolio and to addressing the Grand Challenges in BSS.

This CNSB would receive large-scale proposals and enable data generation and data management and analysis, allowing data integration where scientific activities would be distributed according to expertise. Such a network would accelerate and expand functional understanding due to the scalability of working as a coordinated network. There are several key features of the vision for the CNSB:

- **Coordination.** As the CNSB will span geographic, scientific, and Office of Science boundaries, we recommend the creation of a committee to oversee and manage these activities. The committee should comprise representatives from each of the User Facilities, as well as members of the scientific user community. The committee should meet monthly via video conference and quarterly in person in order to get this ambitious project up and running.
- **Virtual Platform.** The CNSB must have a common platform for proposal, data, and identity management. The various User Facilities all have deployed systems to solve these various issues. However, to be successful at scaling cross-facility efforts, a common system is needed. These features potentially could be built as an extension to or on top of the DOE KBase project. KBase is a powerful virtual platform for analysis that is backed by high-performance computing resources across the DOE ASCR facilities.
- **Communications, User Engagement, and Outreach.** The proposed multisite User Facility network will need coordinated outreach and user engagement activities to attract users to take advantage of the network of facilities. It is recommended that new models be explored for user engagement/service such as the Energy i Corps (innovation Corps) that leverages tools from design thinking to encourage facilities to directly engage their user communities. A communications package should be developed to highlight the opportunities and advantages of

working with the network, consisting of both written and electronic materials. A website could be developed to provide a central hub for information on the network and propagated through the use of social media.

Recommendation

2.31 Establish a Coordinated Network for Systems Biology, a multisite network comprising existing BER and other DOE User Facilities that coordinates multiomics approaches performed with broad spatial and temporal scales to address large-scale and complex challenges for understanding biological systems.

Chapter 3. Earth and Environmental Systems

CHARGE 1 RESPONSE

Alignment of User Facilities to Current Earth and Environmental Systems Research

BER User Facilities are well designed to examine processes that underpin environmental systems. Process-level observations and studies support conceptual understanding of highly coupled systems, which ultimately must be represented in environmental and Earth system models (ESMs). For example, through its combination of fixed and mobile assets, the Atmospheric Radiation Measurement (ARM) User Facility has sampled climate processes in approximately 30 different environments, each with unique properties. These numerous deployments have collectively helped to build a representative sample set of the variability that occurs spatially across the Earth system, providing foundational information that is needed for developing ESMs.

Flexibility of application is another asset for User Facilities, particularly in response to rapidly changing environments or catastrophic events. For example, AmeriFlux and the National Ecological Observatory Network (NEON) have rapid response deployable assets including mobile deployment platforms and towers that can target impactful events like hurricanes, wildfires, volcanic eruptions, and others. DOE's Environmental Molecular Sciences Laboratory (EMSL) also has a rapid science user proposal process that can enable the facility to make measurements at opportune locations and times in response to user-driven priorities. EMSL, in particular, prides itself on providing a problem-solving environment and customizable workflows that can serve evolving scientific needs. This adaptability is a strength that could be expanded through other facilities.

In many cases the User Facilities have worked to remain at the cutting edge of technology by incorporating new, state-of-the-science instruments and approaches. Within the ARM facility, technological advances include employing the newest in ground-based remote-sensing systems, such as polarimetric radars and multiwavelength LiDARs (Light Detection and Ranging method for remote sensing), to observe detailed properties of atmospheric hydrometeors. ARM has also invested in unmanned aircraft systems that will enable new and expansive approaches to observing atmosphere and surface phenomena that previously have been out of reach to classical measurement systems. EMSL has invested in numerous relevant areas including scanning electron microscopy (SEM) and single particle laser ablation time-of-flight mass spectroscopy (SPLAT-MS) to explore the chemical composition of aerosols.

The growth in user base for BER User Facilities is due to more than just the facilities' unique capabilities, but also is largely attributable to the success of BER data models. Large investments across BER in data management, archival, and distribution systems support efficient and broad dissemination of data and results. When data are easily accessible and clearly documented, the information is more widely used. BER is a model for other agencies and institutions on how to confront users with data.

CHARGE 2 RESPONSE

Alignment of User Facilities to Address Future Needs and Grand Challenges in Earth and Environmental Systems

Reaching beyond BER's current research portfolio, the alignment of User Facilities with BER Grand Challenges can provide insight into the role these facilities can and will play in enabling the innovative and high-impact research that is needed to rise to these challenges in coming years. User Facilities have served Earth and Environmental Systems (EES) Grand Challenges relatively well and do provide a strong foundation for further progress and development. In this section, the alignment of User Facilities with each ESS Grand Challenge is briefly assessed. As context, information on User Facility relevancy to each challenge is summarized, based on inputs from the Grand Challenges report itself[8] and more current surveys of the User Facility managers.

Grand Challenge 3.1: Advance Earth system modeling using a hierarchy of models, from process-resolving coupled models to reduced-order models, to transform understanding of the coupled Earth system and to produce useful and credible simulations and predictions of Earth system behavior at multiple timescales.

The ARM User Facility has supported hierarchies of models since the original development of the Cloud-Associated Parameterizations Testbed (CAPT) and associated Single Column Models (SCMs) to test cloud process representations from DOE's ESMs using ARM observations. Both SCMs and the CAPT framework are run deterministically using observed, time-evolving meteorological states to test the fidelity of the simulated cloud systems and precipitation events against those observed and to determine which aspects of the cloud process representations would benefit from further improvement, calibration, or replacement. This strategy is particularly effective since the disagreements between simulated and observed cloud and precipitation fields are dominated by errors in the subgrid physics parameterizations rather than by errors in the fully resolved meteorological states. For example, DOE-supported investigators have used this model hierarchy to understand and reduce the double Inter-Tropical Convergence Zone (ITCZ) bias in the Community Atmosphere Model developed jointly by the National Science Foundation (NSF) and DOE. ARM has recently deployed the large-eddy simulation (LES) ARM Symbiotic Simulation and Observation (LASSO) capability, employing code that is relevant to superparameterized (SP) versions of DOE's Exascale Earth System Model (E3SM). Currently, LASSO is directed toward the ARM Southern Great Plains (SGP) facility. Improvements and enhancements in parameterizations resulting from the confrontation with ARM data can, thereby, be transferred readily to SP-E3SM. SP-E3SM is slated for deployment at Oak Ridge National Laboratory's new Summit Leadership Computing Facility (LCF) to take advantage of Summit's exceptional graphics processing unit (GPU)–based computing capabilities.

At the land surface, AmeriFlux observations have been instrumental in improving E3SM representations of photosynthesis, respiration, surface energy fluxes, and the hydrological cycle. These observations are integrated within the International Land Model Benchmarking (ILAMB) project, which is now being systematically developed for E3SM and Community Earth System land surface model development. Past DOE investment in Free-Air CO_2 Enrichment (FACE) experiments and the current Spruce and Peatland

[8] BERAC. 2017. *Grand Challenges for Biological and Environmental Research: Progress and Future Vision; A Report from the Biological and Environmental Research Advisory Committee*, DOE/SC-0190, BERAC Subcommittee on Grand Research Challenges for Biological and Environmental Research (science.energy.gov/~/media/ber/berac/pdf/Reports/ BERAC-2017-Grand-Challenges-Report.pdf).

Responses Under Changing Environments (SPRUCE) project sponsored by BER's Terrestrial Ecosystem Science (TES) program have provided a mechanistic underpinning of the effects of rising atmospheric carbon dioxide (CO_2) on terrestrial ecosystem function. Given the manifest benefits of model hierarchies for studying physical systems, exploration of such hierarchies for biological systems using the rich array of measurements collected by EMSL and DOE's Joint Genome Institute (JGI) is also warranted.

Grand Challenge 3.2: Establish new observational technologies and use them to understand human and Earth system processes, such as land-atmosphere interactions, biogeochemical cycles, and subsurface soils, to estimate critical process parameters using novel analysis methods, such as machine learning and data science, and to quantify model errors.

User Facilities across BER have a strong history of developing and incorporating new and advanced technologies to address many scientific priorities. In addition to new instruments, it is important to also consider new applications of existing technologies. In recent years ARM has added Doppler wind LiDARS and new arrangements of surface turbulent flux stations to help target turbulent land-atmosphere interactions on multiple scales. These approaches could be applied to new domains (i.e., forest canopy turbulent exchange with the atmosphere). EMSL has developed a very high sensitivity 21-Tesla Fourier transform ion cyclotron resonance mass spectrometer that is enabling a much more detailed view of, for example, soil chemical and elemental cycles. Additionally, new developments around unmanned aerial systems, such as ARM's new Arctic Shark, provide a tremendous opportunity to directly observe processes that were previously out of reach and to help provide spatial context and variability for key processes. To maintain cutting-edge capabilities, each User Facility should have a baseline annual investment in new and adapted technologies, along with specific strategies for their use and reporting on progress.

Grand Challenge 3.3: Advance basic knowledge and scale-aware simulation capability for Earth system feedbacks associated with aerosols and moist processes to better quantify aerosol forcing, precipitation changes, and extreme events with consequences for energy and water cycles, global distribution of nutrients, and human health.

BER and related User Facilities are generally well aligned with the area of "aerosols and moist processes," with numerous facilities targeting specific aspects of this challenge related to advancing basic knowledge. EMSL supports foundational research into aerosol particles, their composition, and their behaviors. ARM provides extensive observations of aerosols in a natural environment and atmospheric processes that drive precipitation. AmeriFlux, Next-Generation Ecosystem Experiments (NGEE), and NEON have focused on interactions at the surface that link precipitation, diffuse and direct incoming radiation, surface turbulent fluxes, hydrology, and ecosystems. Ultimately, the ability to understand the consequences of these climate-relevant processes on large-scale systems requires integration with various modeling tools and, specifically, the ability to develop scale-aware capabilities that cut across relevant scales. Each facility approaches the issue of scale in unique ways. Extensive networks like AmeriFlux and NEON provide measurements at many locations, giving an unparalleled view of large-scale variability, while NGEE has examined spatial heterogeneity at much finer scales albeit in targeted environments like the tropics and arctic tundra. ARM has taken a multiscale approach around its SGP facility that incorporates scanning instruments and distributed installations to cover a variety of scales from individual clouds to storm systems. While the foundation is in place at these facilities, successfully addressing this Grand Challenge will require stronger integration between specific facilities and their relevant modeling systems. Model needs should help to specify how the observational assets are designed and operated, while multiscale observational analyses are then needed to develop and assess scale-aware model parameterizations.

Grand Challenge 3.4: Advance modeling and understanding of important ecological, biological, and carbon cycle interactions and feedbacks in the climate system to identify potential tipping points and possible energy strategies.

Identifying tipping points in Earth and environmental systems requires a certain level of continuity in relevant observational data streams, the context that is granted from having many observations made jointly, and the ability of predictive models to integrate this information and represent emergent behavior. Each facility has its own unique strengths and weaknesses in this regard. On certain temporal and spatial scales, individual BER User Facilities are well positioned to address this challenge, but not on others. For example, there are few time-series data for genomics, and those data that do exist often have limited associated environmental data; without this temporal and environmental context, it is difficult to establish baselines and identify critical transition processes that might indicate tipping points. Furthermore, genomic data are often not sufficiently quantitative for rigorous comparison with environmental or process data, nor for translating what are known to be microbial, metabolic controls over biogeochemical processes to the larger-scale processes they influence, from carbon cycling to climate at the global scale.

It is essential that User Facility measurements be made at the right locations to identify fundamental tipping points. For example, with NGEE-Arctic, the ARM sites on the North Slope of Alaska, EMSL studies of carbon transformation, and other Arctic-centric activities, BER is reasonably well situated to identify tipping points related to thawing permafrost. On the other hand, BER is less well aligned to identify tipping points related to the response of coastal environments to sea level rise.

We note that existing DOE investments in User Facilities have focused on enhancing the scientific community’s ability to make fundamental measurements that inform mechanistic processes and model design. Often, User Facilities have played a key role where the location, complexity, or cost of instrumentation serves as a significant barrier to scientific inquiry by individual principal investigators (PIs). The achievements of BER’s existing User Facilities speak to the success of this conceptual model.

In this context, it is also important to note that a similar class of barrier to scientific progress now exists with respect to use and manipulation of BER’s fully coupled ESM (i.e., E3SM) and its integration with other components of BER’s research program. In part because of increasing software complexity and required knowledge of model architecture, we note that it is extremely difficult for individuals outside the modeling community to execute model simulations for hypothesis testing or the design and exploration of future scenarios of global environmental change. This is especially true for analysis of transient biogeochemical processes that requires many levels of spin-up of ocean and terrestrial ecosystem state variables. As a consequence, many investigators that contribute to BER’s observation and experimental programs are not able to effectively use the tool for site selection, model-data Intercomparison, parameter optimization, or synthesis. We also note that with evolving international agreements and rapid changes in the Earth system, there is also a need to rapidly design and execute the model for new future scenarios that fall outside the domain of well-accepted Intergovernmental Panel on Climate Change (IPCC) concentration pathways. Currently, there are no effective means for “users” to rapidly deploy the model for these purposes.

To this end, it is worth considering whether the User Facility concept should be extended to the simulation domain, to build a broader user base and increase the visibility and impact of BER’s investments in computation and simulation.

Grand Challenge 3.5: Characterize, understand, and model the complex, multiscale water cycle processes in the Earth system including the subsurface to understand and predict water availability and human system response to extremes.

Water cycle processes play out in myriad interactions and feedbacks across the climate system, and, importantly, they are coupled with nearly all environmental systems including human systems. Within BER User Facilities, there are existing capabilities to examine many aspects of the water cycle including precipitation, hydrology, and subsurface processes. This Grand Challenge would be well served by higher-level coordination across facilities and, in some cases, across additional agencies, collaborations which would support better cross-linking of water cycle research. Water isotopes can be a critical factor here, providing key links to climate (through ice cores), hydrological processes within ecosystems, and drainage systems. These measurements, when coupled with other environmental data, can reveal water sources, sinks, and transformations across the Earth system. While human systems are highly coupled with the water cycle, BER User Facilities are not well positioned to characterize the human system response to extremes in drought, flooding, and sea level rise.

Grand Challenge 3.6. Understand the time-dependent processes and mechanisms associated with melting glaciers, ice caps, and ice sheets and their contributions to regional sea level rise.

The melting cryosphere and its impact on sea level represent a complex problem that involves the cryosphere itself; land-ice interactions; ice-ocean interactions; the land-sea interface along coastlines; impacts on biogeochemical cycles; interactions with offshore microbial communities; and many other processes important to climate, climate feedbacks, ecosystems, and human civilizations. During the crafting of the 2017 Grand Challenges report,[9] there were no direct links between User Facilities and this important global challenge. However, based on feedback from the facilities themselves, there is some alignment. For example, the ARM program operated one of its mobile facilities at McMurdo, Antarctica, with some equipment installed on top of the West Antarctic Ice Sheet. Additionally, E3SM has a dedicated priority research theme to examine cryosphere-ocean changes and their interactions with the climate system. However, even considering these areas of alignment, it is clear that BER User Facilities are not well positioned to address either the “time-dependent processes,” which would take a much longer observational effort, or many of the essential “mechanisms,” some of which are currently outside of the BER User Facility capabilities. For example, processes related to subsurface interactions of ice sheets with land and ocean are lacking. Other agencies, such as the National Aeronautics and Space Administration (NASA), may be in a better position to address key aspects of this challenge, and BER should build partnerships where possible. Additionally, BER could consider adapting or developing capabilities to target specific aspects of the challenge, such as surface energy budgets, precipitation, and other processes.

[9] BERAC. 2017. *Grand Challenges for Biological and Environmental Research: Progress and Future Vision; A Report from the Biological and Environmental Research Advisory Committee*, DOE/SC-0190, BERAC Subcommittee on Grand Research Challenges for Biological and Environmental Research (science.energy.gov/~/media/ber/berac/pdf/Reports/ BERAC-2017-Grand-Challenges-Report.pdf).

Grand Challenge 3.7: Quantify the interplay between internally generated climate variability and externally forced response involving anthropogenic and natural factors and their relative roles in the time evolution of regional variability to understand predictability of the Earth system.

While robust attribution of climatic fluctuations to natural versus anthropogenic factors is still an aspirational research objective, ARM has enabled several path-breaking studies detecting and isolating external (anthropogenic) forcing by greenhouse gases from the much larger fluctuations in the Earth's radiation fields due to natural thermodynamic variability. The signals of both CO_2 and methane (CH_4), the first and second most-important anthropogenic greenhouse gases, have now been detected in the ARM dataset. The distinctive properties of the ARM User Facilities that enabled these studies include the collection of the full suite of radiative, thermodynamic, and chemical variables required for robust detection over timescales of decades, together with stable and ultra-accurate calibration of the critical infrared radiance interferometers. The ARM investigators also quantified how these greenhouse signals are strongly modulated by natural seasonal cycles in both temperature and water vapor.

AmeriFlux similarly has played an important role in examining long-term multidecadal trends in surface evapotranspiration as well as impacts of El Niño–induced drought and other forms of climate variability on the terrestrial carbon cycle. With many sites now exceeding a +15-year time series, the network is well poised to study the effects of longer-term variability in climate system on terrestrial ecosystem function and feedbacks.

The EMSL and JGI User Facilities have supported comprehensive studies of the greenhouse gases released by the microbial communities in melting permafrost. The core objective is to understand how much of the carbon stored in the originally frozen permafrost is expressed as CO_2 or as CH_4, which has a global warming potential 32 times higher than that of CO_2 on less than centennial timescales. The combination of genetic sequences from JGI enables a thorough characterization of the microbial populations in the permafrost samples, and the proteomics and transcriptomics provided by EMSL reveal the metabolic functions of these communities that result in greenhouse gas emissions.

Grand Challenge 3.8: Understand the long-term Earth system stability in response to possible future Earth system outcomes and address the level of confidence and identify emergent constraints for the range of model projections.

The ability to understand, constrain, and quantify long-term Earth system stability inherently requires robust ESMs that represent relevant climate processes and feedbacks. Long-term and process-based observations must be the foundation for creative model assessment to ensure the quality of long-term simulations. While the 2017 Grand Challenges report did not identify any alignment between BER User Facilities and this specific challenge, numerous facility managers did identify alignment. Some specific areas of alignment are related to terrestrial carbon sinks and sources, biogeochemical cycles, and long-term examination of ecosystem stability in the face of environmental change, but these could be linked more holistically, including a focus on anthropogenic factors. Overall, making progress on this challenge requires directed input from Earth system modelers to ensure that User Facilities are targeting the essential processes. Additionally, mining past data is needed to better establish baselines for reference when considering tipping points and emergent constraints. Addressing this Grand Challenge also suggests BER should explore new means to promote synthesis and analysis across the different User Facilities and PI-driven research programs.

Crosscutting Opportunities for User Facilities

Given the overall assessment of alignment with Grand Challenges, the User Facilities face a number of "crosscutting" opportunities to address evolving and cutting-edge research priorities. These opportunities are briefly outlined below, while concrete actionable plans or concepts for meeting these opportunities are delineated in subsequent sections.

Integrative Science

While great progress has been made in the past couple of decades toward understanding and modeling many disciplinary aspects of Earth and environmental systems, the state of the science is moving largely toward a more integrative approach that examines coupling across physical, biological, chemical, and human systems. Although this interdisciplinary coupling is the new frontier in model development, making significant progress toward developing these models requires new approaches to observing processes that cut across systems. One example is the terrestrial-aquatic interface (TAI), wherein there exist important links between biogeochemical processes and hydrology, among others, but there are no long-term observing sites for this interface. The BER Grand Challenges touch on a number of other key areas that require new observational approaches, including the water cycle and interactions with the subsurface environment, biological roles in the carbon cycle, ice sheet–ocean interactions, and many others. Addressing these challenges and probing the relevant interdisciplinary processes require new approaches to linking and leveraging BER User Facilities, both together and in coordination with other networks like AmeriFlux and NSF's Long-Term Ecological Research Network (LTER). In many cases, a process must be established whereby large, crosscutting science challenges can be met with coordinated resources from multiple facilities.

Scaling

BER's scientific portfolio and interests span a continuum of scales ranging from nanometer to global, with essential linkages across many of these scales. In many cases, understanding the intrinsic behavior of a system at microscales clarifies the manifestation of that system at a macroscale, and vice versa. Scaling is particularly important when considering the interface between observations and ESMs. User Facilities might provide very detailed observations of a cloud field or aerosol composition, but clouds and aerosols must be represented globally in ESMs. Thus, it is essential that the context for measurements from User Facilities is well understood (i.e., where they fit within the continuum of relevant scales) and that those measurements themselves are designed in a way to enable up- and down-scaling as appropriate.

Adaptability

As embodied by many of the BER Grand Challenges, scientific frontiers continue to evolve to target emergent processes in the climate system, advances in technological capabilities, increases in computational power, and continual advancement in knowledge of the Earth system. User Facilities must remain agile and adaptable enough to respond to changing scientific priorities in a timely, and sometimes rapid, fashion. Examples might be a rapid destabilization of an ice sheet, terrestrial ecosystem response to changes in hydrology or drought, or breakthrough technological advances in the ability to probe a biome in much more detail. User Facilities need a process to identify and respond to rapidly evolving scientific needs and priorities. This adaptability might include reserving some fraction of resources for focusing on new "cutting-edge" research or adding a rapid response capability where appropriate.

Linking Data and Knowledge with Stakeholders

Without doubt, the BER User Facilities provide a strong foundation of process understanding for Earth and environmental systems, with capabilities that in many cases are unmatched globally. The challenge rests in how best to transfer that process-level knowledge to diverse and impactful applications such as operational utilities, predictive models, satellite assessment, and others. For all facilities there are additional opportunities to consider effective means of packaging and synthesizing data for target audiences. Additionally, there is large potential to better combine information across User Facilities both within and external to BER to serve broader user communities and to enable advances. For example, data from multiple facilities relevant to the theme of hydrological cycles could be synthesized in a way that is more effective for ESM development. This synthesis of information can also extend to model data itself, where observation-model synergies can lead to advances in process understanding beyond what is possible from observations alone.

Supporting Facility Science

To enable the most effective use of User Facilities to achieve priority BER scientific outcomes, it is essential to consider the relationship between the use of "technical" resources (i.e., the User Facilities themselves) and "science" resources, which involve the research conducted by investigators based on the technical resources. In some cases, there is a disconnect between the funding mechanisms for these two sets of resources. For example, ARM projects are supported without directly associated funding for the proposing investigators. A similar challenge exists when attempting interagency collaboration around BER User Facilities, where memoranda-of-understanding or some other approach might help to clarify and solidify the relationships between Facilities and users. Additionally, there seems to be wide agreement across the scientific community that there is a wealth of information produced by User Facilities that is never analyzed sufficiently. Opportunities exist to support additional scientific research through stronger engagement with scientific funding programs either within BER or at other agencies.

CHARGE 3 RESPONSE

Development of Additional User Facility Capabilities

Existing challenges and opportunities for addressing high-priority EES research motivate the need for new User Facility capabilities, both within existing facilities and through the development of entirely new facilities. The aim of expanding capabilities is to ensure that User Facilities (1) effectively adjust to evolving requirements driven by the research community and Grand Challenges and (2) remain flexible enough to continue enabling the cutting-edge research of the future. Concrete ideas for new capabilities have been drawn from numerous sources, including the BERAC User Facility assessment workshop, the Grand Challenges report itself, and other thematically focused BER workshops. Each proposed expansion is contextualized using a short narrative, followed by concrete recommendation(s).

Guiding User Facility Instrumentation and Operations

The User Facilities serve and operate within a broader context of community-driven research agendas, and thus the instruments and capabilities of the User Facilities must advance to support current evolving research needs but also to anticipate near-term future research directions. Due to the diversity of research priorities addressed by these facilities and the highly heterogeneous approaches to the priorities being pursued at any one time, a wide range of measurements is inherently required for all these approaches. Therefore, it is evident that the User Facilities have to prioritize the acquisition and deployment of promising new technologies based in part on maximizing their value-added benefits to the widest possible sectors of the user community. This prioritization will by necessity involve estimating both the scientific return per unit of investment and ongoing support costs as well as the scope and diversity of the scientific community that would realize the benefits. Naturally, other collective scientific facilities face similar tradeoffs, and several have adopted quite rigorous and transparent mechanisms to resolve these issues through open, thoroughly documented processes with ground rules established at the beginning of the evaluation process.

One approach to assign initial estimates of the relative priority would be to consider the pursuit of each of the BER Grand Challenges by means of an experimental program using the User Facilities. Suppose each of the Grand Challenges were accompanied by a small ensemble of science questions, each of which could be converted to a set of one or more falsifiable hypotheses. In turn, the experiments designed to resolve these hypotheses would require collection of observations both from existing instruments as well as potential future measurements. The collection of the latter set of data presumably would determine major new instrumentation and desired capabilities. A critical component of an implementation plan for BER Grand Challenges linked to the User Facilities would consist of a much more detailed roadmap linking individual challenges to science questions and accompanying falsifiable hypotheses, followed by the experiments, required observations, and new instruments needed to collect all the required data streams.

Recommendations

3.1 Develop new technologies that address persistent scientific needs for ARM, including convective vertical velocity, aerosol profiles, ice nucleation, and continuous thermodynamic profiling. Technologies warranting investment to meet these needs include unmanned aircraft systems and tethered balloon instrumentation and miniaturization to access previously inaccessible domains.

3.2 Employ targeted calls for User Facilities to better address specific Grand Challenges with a focus on cross-disciplinary and coupled system studies.

3.3 Consider the mechanisms used by other user communities (e.g., astronomers and high-energy experimental particle physicists) to evaluate and select from candidate augmentations to existing User Facilities. The mechanisms should start from predefined evaluation metrics and definitions of success and should operate in an open and transparent manner with extensive documentation of the prioritization procedures.

3.4 Employ targeted calls for the design of several new User Facilities to better address specific Grand Challenges and emerging research frontiers.

Bridging Scales

Multiscale interactions, and the ability to up- or down-scale information, are essential to many of the Grand Challenges. In some cases, the most important (or most uncertain) scales along the continuum may be changing, for example, as a result of past research, regime shifts, climate change, and/or new discoveries. In addition, the essential or emergent parameters within a system that serve to fundamentally characterize that system and drive multiscale system dynamics must be identified and understood to build a foundation for bridging scales within modeling systems. Often, these emergent parameters are the ones that enable models to simplify complex natural systems. Each User Facility is most directly relevant to a specific range of scales and emergent parameters, and it is essential that each facility has a process for identifying and assessing its capabilities relative to the most important scales and parameters for current and future research. Such identification can be supported using Observing System Sensitivity Experiments (OSSEs), controlled manipulation experiments, or other approaches. Advancing understanding of the coupled Earth system (Grand Challenge 3.1) will require application of observations and models that span the relevant scales and parameters to support development of scale-aware model parameterizations.

As an example, the ARM facility provides detailed observations of clouds and cloud populations at spatial scales ranging from hundreds of meters to a few kilometers, while cloud systems may span spatial scales of hundreds to thousands of kilometers. A strategy to tackle this problem involves some enhancements and directed application of the ARM facility in combination with other existing observation systems. Information about the large scale can be provided by satellite observations and by ground-based networks such as the National Oceanic and Atmospheric Administration's (NOAA) Next-Generation Radar (NEXRAD) network. In a region where there is a distinct gradient in cloud characteristics, such as the central United States, several ARM facilities could be deployed in conjunction with extended ARM facilities with reduced capabilities for sampling thermodynamic and dynamic profiles. Four such profiling systems are currently deployed around the SGP site in support of the LASSO system. These observations should be complemented by model simulations at multiple scales with LES-scale simulations run over the ARM observatories, nested within moderate-scale simulations over the larger experiment domain.

Such a system would provide detailed information at high resolution along with extensive information at the larger scale. An important goal of a multiscale observation and modeling experiment would be to identify emergent characteristics at each scale. As an example, the distribution and timing (with respect to the diurnal cycle) of precipitation along a transect across the central United States are quasistationary and can be captured well by large-scale observing systems. The ARM observations and high-resolution simulations would provide detailed information about local environmental conditions and cloud properties at selected points across the domain. By selecting points that span the conditions of the larger domain, an important research goal would be to determine the fine-scale parameters that determine the large-scale emergent structures.

With the emergence of the theme of microbes to global Earth system function and the related chapter in the 2017 Grand Challenges Report, another opportunity for making progress in bridging scales may arise at the interfaces among JGI, EMSL, and AmeriFlux. Specifically, the collection of continuous genome and proteome time series at a few AmeriFlux supersites may allow for new insight regarding microbial controls on ecosystem function. Fine-scale habitat characterization and resource tracking through EMSL capacities may be key for translating microbial to ecosystem controls. For example, AmeriFlux observations provide evidence for rapid mobilization of microbial communities in the hours and days following rain events that yield a heterotrophic respiration "pulse." Quantifying the temporal evolution of changes in plant and microbial community composition and gene expression during these events, as well as resource and habitat characteristics, may provide new information about microbial controls on soil respiration, nitrous oxide emissions, and CH_4 fluxes. Similarly, coupling time-series information on plant and microbe genomics and proteomics with drought and flooding events, heatwaves, and even more fundamental phenological cycles is likely to yield new understanding relevant for building next-generation models of terrestrial ecosystems and the land surface.

Scaling from the footprint of an AmeriFlux tower to the size of a typical ESM grid cell represents another important scaling challenge that focuses on the role of forest demographics in regulating the carbon balance of ecosystems. While recent satellite and aircraft remote-sensing integration of LiDAR and hyperspectral information have yielded recent breakthroughs in our ability to measure forest biomass at scales of meters to kilometers, we still do not have robust tools for remotely tracking forest canopy dynamics, including recruitment and mortality of trees, changes in the 3D structure of canopies, and other structural changes that influence the terrestrial carbon sink. This information is essential for informing new ecosystem demography models such as FATES. Remote sensing of fine-scale forest structure, we believe, represents a gap between U.S. funding agencies and is well suited for an initiative led by new or existing DOE User Facilities. Specifically, NASA is not moving aggressively to build time-series information at this scale, as reported in the recent National Research Council (NRC) Decadal Survey.[10] The U.S. Forest Service (USFS) would benefit from new technologies to provide this information, but it does not have the necessary research budget to explore needed imaging and drone technologies that potentially could revolutionize its Forest Inventory Analysis (FIA), which currently stands as the backbone of the USFS forest sampling program.

Recommendations

3.5 Identify and determine for each User Facility the controlling emergent processes and behaviors of a system (e.g., the "rare biosphere") at different scales as a means to better constrain how these processes interact across scales and to prioritize facility activities.

3.6 Augment and align ARM resources along a central U.S. transect to better address both small- and large-scale processes associated with the full life cycle of convective precipitation across the central United States. Such a transect could include new sites in Colorado, SGP, and the southeastern United States, with smaller sites in between, links to other networks, and integration with other key environmental transitions (e.g., forest coverage, drought, ecosystem processes, and carbon cycle).

3.7 Develop and employ an appropriate cross-scale modeling framework for each primary User Facility as an instrument to support up- and down-scaling between observations and large-scale models (e.g., LASSO for ARM).

3.8 Employ advanced unmanned aircraft systems in a systematic approach to bridge across scales and to assess the spatial representativity of User Facility observations in a variety of multiscale environments.

3.9 Develop a network of AmeriFlux omics-to-ecosystems supersites, where high–temporal resolution field and laboratory observations of omics, microhabitat-scale conditions, and fluctuating resources are generated automatically and data are compared with ecosystem flux observations and models.

3.10 Build new capacity through a combination of AmeriFlux and ARM technologies to map individual tree structure and seasonal and interannual forest dynamics across the network.

Science Priority: Ice Nucleation

Cloud-aerosol interactions continue to be one of the largest sources of uncertainties in ESMs, and the process of ice particle nucleation is one of the most challenging aspects of this science domain. The study of ice nucleation processes is hindered by the great challenge of measuring ice nucleating particles (INP). Although, in recent years, there has been progress in these measurements, clearly, measurements are still limited to a subset of nucleation modes and sometimes are not sensitive enough to characterize INP populations in their natural state. As a result, much is still not understood about INP concentrations in natural environments and what makes individual particles effective in forming ice. There is only nascent understanding of the distribution of INP particles, the processes through which the properties change in space and time, how these aging processes ultimately impact cloud activity, and other considerations that

[10] National Research Council. 2017: http://sites.nationalacademies.org/DEPS/ESAS2017/index.htm.

can have a profound effect on cloud properties. Ice nucleation is important globally, performing particularly important roles in polar regions, where prevalent low-level, mixed-phase clouds play a key role in a rapidly changing regional system and at much lower latitudes. In these regions, cloud ice formation impacts the vertical distribution of moisture, formation of precipitation, and spatial distribution of upper-level clouds. Broadly, ice nucleation is an area in which a concerted focus is needed to achieve transformational advances in understanding.

Physical and chemical properties of individual particles ultimately determine their propensity to support ice nucleation, likely determining the specific nucleation mechanisms in action. Consequently, detailed and comprehensive analysis of INP at scales ranging from individual particle behavior to the bulk response of particle populations over model–grid box domains are greatly needed to promote significant advancement on this theme. This challenge presents a unique and collaborative opportunity to harness the strengths of ARM, EMSL, and possibly JGI facilities toward a common, high-impact goal. ARM operates two sites on the northern coast of Alaska, one in a largely clean environment and one in an environment heavily influenced by industrial activities in the Prudhoe Bay oil fields. Additionally, ARM has the ability to fly manned aircraft across this Arctic domain. These sites and facilities provide an opportunity to collect aerosol samples on an on-going or targeted basis in a region where INPs play a strong role in controlling the radiative properties of clouds. To enable such measurements would require application, adaptation, or development of new instruments that are sensitive enough and able to operate for extended periods of time. At the same time, the extensive analytical capabilities available at the EMSL facilities provide an opportunity to obtain detailed information about the composition and behavior of INPs. To support these studies, a valuable enhancement to EMSL capabilities would be (1) the addition of a cloud chamber, which permits the observation of INP activity in a realistic but controlled setting, and (2) the ability to examine environmental factors influencing ice nucleation. Such laboratory analysis could be further extended to engage JGI since biological material can act as an INP. Lastly, BER modeling programs and capabilities (e.g., E3SM) can all be brought to bear on this important issue by examining the roles and implications of ice nucleation processes at different scales to assess model climate sensitivities to the specification of ice nucleation modes and, ultimately, to incorporate and test new approaches to ice nucleation.

Recommendations

3.11 Establish a joint facility activity among EMSL, JGI, and ARM, perhaps by extending existing Facilities Integrating Collaborations for User Science (FICUS) collaborations, to develop and implement a comprehensive observational strategy (field and laboratory) to measure and discern modes of ice nucleation under real atmospheric conditions.

3.12 Develop a cloud chamber with the ability to examine aerosol particle formation and cloud activity, with links to EMSL for characterization of organic INPs formed through (photo)-chemical processing of organic precursor emissions.

Science Priority: Cryospheric Change and Sea Level

One of the clearest fingerprints of the changing Earth system is the observed cryospheric change, embodied by rapidly declining glaciers and ice sheets. Cryospheric change, specifically melt, is progressing asymmetrically across the globe, but the influences of sea level can be felt globally through coastal processes and feedbacks at the interface between cryosphere and ocean. Rising seas will have dramatic impacts on human systems, coastal processes, ecosystems, and energy stability. Sea level rise is driven in part by ocean thermal expansion and in part by melting land-based ice. Over the past few decades, glacier melt has dominated this land-ice contribution, but this source has been rapidly overcome by the accelerating contribution from Greenland over the past few years. Antarctic mass loss remains a smaller contribution to sea level rise, but it could quickly become dominant, depending on the stability of the West Antarctic Ice Sheet. Importantly, these cryospheric contributions to sea level rise are dependent on the balance of mass loss (e.g., melt, ice dynamics, and calving) and mass gain through precipitation. Models likely do not correctly represent the mass budget for ice sheets and glaciers, which are not the key drivers of the mass budget.

The E3SM community is distinctly interested in using variable mesh approaches to address the relationships between cryospheric change and sea level. However, BER User Facilities and capabilities generally are not well aligned to provide the observational foundation for relevant model assessment and development. ARM has made observations in Antarctica to provide an initial perspective on some atmosphere processes as they relate to West Antarctica. Given insightful deployment of ARM facilities, BER could contribute significantly in the future toward filling substantial knowledge gaps surrounding the atmospheric drivers of ice sheet melt and the important mass contributions from precipitation over key ice sheets. Addressing the issue of sea level rise more holistically will also require BER to develop advanced understanding of ice sheet dynamics, ice-ocean interactions, and other key processes to which User Facilities are not currently aligned, either via expansion of capabilities or interagency collaboration.

Recommendations

3.13 Deploy the ARM Mobile Facility No. 3 for extended operations at a location relevant for addressing cryosphere impacts on sea level, such as West Antarctica or Southern Greenland.

3.14 Hold a targeted workshop to explicitly consider how BER facilities can address cryospheric change.

Science Priority: Response of Terrestrial-Aquatic Interface to Cryosphere-Driven Hydrological Change

The TAI represents one of the most dynamic regions on Earth that is subject to unique forcings driven by global climate change. The 2016 DOE workshop, Research Priorities to Incorporate Terrestrial-Aquatic Interfaces in Earth System Models,[11] provides a robust plan of action to address the challenges of TAI interface science in tropical and temperate habitats across a range of scales. However, a particularly unique aspect of TAI science was not covered by the report: exploring and documenting the impact(s) of ice melt on TAI biogeochemistry and ecology.

The cryosphere is broadly defined as the Earth's reservoir of frozen water—ice, snow, and permafrost. The cryosphere is broadly distributed from polar regions to alpine glaciers in temperate and tropical regions. The stability of the cryosphere is a function of local and regional patterns of temperature, precipitation (snowfall), and wind. Degradation (loss) of the cryosphere alters local hydrological regimes and may generate significant alterations in TAI processes and dynamics. As the Earth warms, the area of the cryosphere shrinks. As the cryosphere shrinks, the discharge of meltwater and the flux of cryosphere-derived components to adjacent environments increase. Additionally, sea level rise driven by melting

[11] U.S. DOE. 2017. *Research Priorities to Incorporate Terrestrial-Aquatic Interfaces in Earth System Models: Workshop Report*, DOE/SC-0187, U.S. Department of Energy Office of Science (tes.science.energy.gov/files/TAI_Workshop2016.pdf).

cryosphere can strongly impact coastal aquatic zones. The consequences of these hydrological changes on TAI processes must be explored, constrained, and modeled. Potential impacts include alterations in ecological dynamics and in biogeochemical cycling, particularly in terms of carbon storage and remobilization and nutrient cycles. Thus, this theme represents a potential coupling of processes across physical, chemical, biogeochemical, and ecological systems that operate at the interfaces of land, inland waters, wetlands, estuaries, and the ocean.

Together, ARM, JGI, and EMSL are uniquely positioned to advance science in this area, and, given the interdisciplinary nature of the question, the potential for collaboration is significant. The ARM site on the North slope of Alaska may already provide some data that can be used for this science priority. Moreover, ARM mobile facilities could be operated in targeted coastal or watershed-based terrestrial-aquatic zones of particular relevance to, and alignment with, ecosystem changes and biogeochemical cycles. The analytical capabilities of the EMSL facility provide a unique means to assess changes in biogeochemical flows, including tracking mobilized carbon and nutrients as well as microorganisms (identity and activity) that consume cryosphere-derived materials. Facilities at JGI can provide detailed information on how biological communities are altered by cryosphere destabilization.

Recommendation

3.15 Hold a targeted workshop that builds on the prior TAI workshop, broadening the scope to include areas further from the coastlines in both directions as well as the impacts of the changing cryosphere. Workshop outcome: Framework for how User Facilities can address evolving needs on this theme.

Science Priority: Enabling Manipulation Experiments

Manipulation experiments permit assessment of known factors that drive ecosystem dynamics and, in many cases, reveal unknown and unanticipated regulatory factors, feedbacks, or networks. Field and laboratory manipulation experiments are challenging and require robust controls, statistical power (more than adequate replication), and a holistic suite of analytical variables and parameters. Timescales are also a key consideration. The complete suite of impacts resulting from an experimental manipulation may require long periods of time, in some cases years, to be fully expressed and revealed. Relevant timescales include those that capture physiological responses, those that help identify ecological responses, and those that constrain evolutionary responses. Reductionist approaches lead to a mechanistic understanding of the dynamics.

Field manipulation experiments are worth pursuing in habitats or environments that are anticipated to endure severe future change (e.g., exploring drought or fire effects in inland grasslands), that could serve as a strong model for understanding change across scales (e.g., an experiment positioned to scale from genes to ecosystems), or that are situated at an ecosystem interface (e.g., the TAI). Field experiments like SPRUCE are the gold standard for such undertakings because of the spatial scale of the experiment, the comprehensive suite of measurements and participants, and the extensive instrumentation that supports data-model synthesis and forecasting. Field observations such as NGEE-Arctic and NGEE-Tropics are also extremely valuable. There is also a continued role for controlled experiments in mesocosms, or ecotrons (worth noting is the French example: www.ecotron.cnrs.fr), which are useful in that they reduce the complexity of the natural world. Manipulations designed to assess biogeochemical dynamics and responses and/or to assess the impact of biodiversity on the ability to endure stressors may be more easily carried out in ecotrons.

Addressing this challenge requires investment in experimental systems and instrumentation. Examples are quantification of carbon and nitrogen gas fluxes as well as evapotranspiration; environmental conditions (meteorology); and vegetation, soil, and microbial community characterization. This work could foster collaborations between JGI and EMSL, and there may be a role for ARM. Such a program would add to the already broad suite of instrumentation and expertise available at EMSL and help to bring EMSL to the field. Close ties with JGI will provide parallel assessment of biological, especially microbial, community dynamics. These studies will help address the call for action in the 2017 Grand Challenges report: to "conduct experiments that help determine the influence of microbial processes at larger, aggregate scales."[12]

Recommendations

3.16 Develop a framework that leverages the full suite of capabilities at EMSL and JGI to conduct manipulative experiments using ecotrons.

3.17 Consider a targeted research announcement aimed at supporting manipulative field experiments that leverage EMSL and JGI collaborations.

3.18 Establish a User Facility to enable manipulative experiments at field-relevant scales that are critical for advancing our understanding of the linkages between physical and biological systems and across scales of organization, from molecules to habitats to ecosystems.

Advancing Data Analysis Capacity

The exponentially increasing capabilities of measurement systems, in particular those associated with imaging and with the analyses of biological samples, require a concerted and coordinated response to ensure the data recorded by BER's User Facilities can be used effectively to address Grand Challenge questions. For example, the cumulative number of sequenced human genomes doubles every 7 months. This is just 40% of the 18 months needed to double transistor counts, known as Moore's law, and 50% of the time for total available supercomputer floating point operations per second to double worldwide. At the same time, Kryger's law, which had predicted that the cost of storing this data would halve as fast or faster than transistor counts would double, has ceased to apply as the economies of scale have been wrung out of existing disk technologies. In short, the rates of DNA sequencing are growing exponentially faster than the resulting sequences can be either analyzed or stored. Currently, 159 billion bases are sequenced per day at JGI, posing an increasingly impracticable present and future challenge for the analysis of the output from just this single BER User Facility.

The implications of these trends are that the amount of data per unit mass of biological sample will continue to grow exponentially for the foreseeable future. For this data to be useful, BER's JGI and EMSL User Facilities will need to make conscious investments to support the end-to-end analytical workflow and will need to borrow from other disciplines, particularly those in DOE's Office of Advanced Scientific Computing Research (ASCR), that are exploring novel data sampling and data reduction methods. These include just archiving data at the point of collection using either traditional rule-based algorithms or classification techniques using machine learning. For example, for the Large Hadron Collider (LHC), this

[12] BERAC. 2017. *Grand Challenges for Biological and Environmental Research: Progress and Future Vision; A Report from the Biological and Environmental Research Advisory Committee*, DOE/SC-0190, BERAC Subcommittee on Grand Research Challenges for Biological and Environmental Research (science.energy.gov/~/media/ber/berac/pdf/Reports/ BERAC-2017-Grand-Challenges-Report.pdf).

strategy results in orders-of-magnitude reduction in the proton collisions that are subjected to subsequent thorough analysis. The novel methods also include advanced “lossy” data-compression techniques originally developed for commercial image processing that can yield an order-of-magnitude compression, and increasingly machine learning is also being explored for its data-compression potential.

The workshop participants recommend that the standard should be that a unit of experimental sample (i.e., a gram of biological material) can be analyzed and the data archived and made available in a fixed unit of time, despite the exponentially larger amount of data being extracted from that sample. This is analogous to the standard adopted by the climate modeling community that each new ESM generation be capable of simulating 5 years of climate per “wall-clock” day, despite its increased resolution and multiphysics process complexity. Adoption of equivalent standards at BER’s User Facilities would help ensure that these facilities deliver results in a timely way to maintain the active engagements of their user communities.

Recommendation

3.19 Encourage concerted coordination between DOE ASCR and the BER User Facilities to improve the pace of data archival, the quality of metadata, the ease of data access, and tools for data analysis.

Tools for Synthesis

The detailed state-of-the-art measurements made by DOE User Facilities often require special expertise in measurement techniques, uncertainties, and interpretation, making it difficult for nonspecialists to use the data generated from the measurements. This issue is compounded when a suite of diverse measurements is needed, as is often the case for Earth system modeling studies. The models themselves have the same challenge—scientists with the requisite expertise to run models like E3SM are a rarified group, yet there are many more scientists in the BER community with strong interest in testing and applying existing models and in helping develop new ones, especially those that bridge scales. The democratization of ESMs is an important synthesis goal, because it will broaden the intellectual scope of the community, pushing the science forward. Both workshops, the Grand Challenges Vision and the User Facilities, identified the need for BER to explore ways to enhance support for data and computational synthesis.

To better bridge the gap from measurements to modeling and analysis, there is a need for (1) tools for dataset syntheses that transform detailed measurements to the quantities that are needed for analysis and modeling, (2) improved organization of data for following community standards to enable easier and broader use, and (3) open-source contributions of synthesis tools to leverage community-wide expertise and knowledge. To address these needs, User Facilities should support making measurements more comprehensible to a diverse user base by developing standardized data products, tools for synthesis of measurements from multiple sources, and facilitation of “crowd sourcing” models (e.g., GitHub) to leverage community expertise and interest in User Facility measurements.

Within BER there are existing efforts toward this type of synthesis. The ARM facility supports software development in support of value-added data products (VAPs), which convert the detailed measurements taken by the ARM instrument suite to geophysical quantities for use in scientific analysis. The ARM Best Estimate (ARMBE) group of products further synthesizes diverse VAPs onto a common time and height grid compatible with global climate models aimed specifically toward interfacing the observations with ESM development activities. The ARM facility also hosts a GitHub repository where community members can contribute and further develop existing codes for product generation or analysis. The Python ARM

Radar Toolkit (PyART) is one example of a successful open-source framework for sharing and developing useful software tools. Excellent examples of data synthesis activities also exist within the larger community. Specifically, the World Climate Research Program governs the Observations for Model Intercomparisons Project (Obs4MIPs) activity, which aims to make observational data products more accessible for the intercomparison of ESM simulations. These observations and others are integrated within ILAMB, which is supported by BER and has been used extensively in recent work for E3SM and the Community Earth System Model (CESM) land surface model development. Data accessibility is accomplished through the definition of data and metadata standards, critical datasets, standards for inclusion, accompanying documentation, and feedback among modeling participants. BER User Facilities should build on these existing activities toward more comprehensive and coordinated synthesis tools that bridge the relevant scales and disciplines, improving accessibility and applicability of the DOE User Facility measurements.

Several specific activities have been identified to accelerate the development and application of tools for the synthesis of User Facility measurements. Automated production of synthesized data products, developed with direction from users involved in analysis and modeling studies, should be continued and expanded. A particular focus of this activity should be to build stronger links with complementary data from other agencies. Specific examples include linking satellite observations at overflight times and output from reanalysis datasets to measurements on the ground, such as from the ARM, AmeriFlux, NEON, and LTER sites. Another pathway to facilitate the use of User Facility observations in modeling studies involves the use and development of forward operators, or "instrument simulators." Simulators are algorithms that use model variables to compute instrument observations for model-observation intercomparison. Ecological forecasting through efforts such as the Ecopad at the SPRUCE site is another example, where real-time data collection is compared to and informs an ecosystem-scale model. Such approaches have been used successfully for evaluation of ESMs using active and passive remote-sensing observations. Development of such approaches should be considered and expanded beyond current efforts, applying them to more measurement types and User Facilities. Overall, the use of cutting-edge software development techniques and technologies for developing synthesis data products will increase the engagement of a larger research community, particularly researchers involved in modeling and analysis, increasing the impact of User Facility measurements and progress toward addressing BER Grand Challenges.

To integrate across theory, models, and observations, there are needs to (1) enable a broader community of users to work with the most complex and state-of-the art models, (2) establish a platform where multidisciplinary teams of scientists can develop, evaluate, and integrate models across scales (from omics to the Earth system), and (3) broaden capacity for model-data synthesis. BER also needs model-data synthesis capacity, including experts in model development, simulation, and analytics; dedicated computational resources to enable data availability, analytics, and synthesis across BER disciplines; coordination of model evaluation capabilities, model testbeds, and tools; and enhanced capacity to curate fine- to global-scale model results, large observational data, and model-data fusion products.

This goal for synthesis reiterates ideas expressed in the Grand Challenges report, which specifically calls out the need for synthesis via a computational User Facility for "rapid design, generation, evaluation, and diagnosis of ESM simulations" and "data-model synthesis." As data volumes continue to increase exponentially, the EES community recognizes the clear need for new investments in synthesis to more effectively exploit the science content of datasets generated from multiple User Facilities that are constraining different and complementary aspects of the coupled Earth System.

Recommendations

3.20 Develop a living and broadly accessible repository of analysis tools, with collaborative links to the various research programs that support the tools' development and use.

3.21 Consider aggregating tools for data analysis, data-model synthesis, and state-of-the-art simulation modeling into a software container that could be used at users' institutions, on User Facility computational resources, or in the cloud. This would help maximize the range of options and efficiency for analysis of User Facility data and of community models.

3.22 Implement the call in the Grand Challenges report to develop a computational and synthesis User Facility that supports the rapid design, generation, evaluation, and diagnosis of ESMs, including robust data-model synthesis. This facility will support the accessibility and availability of models and simulations to a wide community of potential users; and the development of new models addressing scaling across organization over the full purview of BER (omics to Earth).

CHARGE 4 RESPONSE

Opportunities for Collaboration Among User Facilities

Collaborations in multiple directions will enable BER to best integrate new, and leverage existing, User Facility capabilities toward addressing Grand Challenges and BER strategic research initiatives. For the EES theme specifically, collaborations can enable BER to strengthen interdisciplinary research and bridge across important scales, both of which challenge any single User Facility. Supporting these collaborations requires consideration of the manner in which User Facilities and scientific research are supported within BER, and the ability to attract and train a broad user base that is able to confront new complex science by leveraging User Facility capabilities.

Supporting Interdisciplinary Science

The BER Grand Challenges report outlines a series of recommendations for scientific research, much of which is interdisciplinary and spans across User Facilities. These Grand Challenges should serve as a guide for User Facilities; however, coordination among these facilities is needed to truly achieve interdisciplinary outcomes.

In some cases, the role of different User Facilities in addressing a particular aspect of the Grand Challenges is clear. For example, addressing Grand Challenge 3.3 requires information on aerosol chemical composition and ice nucleation, and it would clearly benefit from a coordinated effort that leverages EMSL, ARM, and JGI. Addressing the land-atmosphere interactions and biogeochemical cycles component of Grand Challenge 3.2 would similarly benefit from deployment of multiple User Facilities at biologically interesting sites (e.g., employing some combination of NGEE, NEON, AmeriFlux, SPRUCE, or ARM). Improved understanding of climate forcings and responses of permafrost thaw pertaining to the carbon cycle is needed for Grand Challenges 3.2, 3.4, and 3.5, and this research would benefit from a coordination between NGEE-Arctic and ARM.

In other cases, while it is clear from the interdisciplinary nature of the problem that multiple User Facilities are needed, understanding which facilities and how they can function effectively together warrants further discussion. In these cases, workshops are needed to develop a vision for how User Facilities can collaborate and coordinate to address the cross-disciplinary themes. The charge for such workshops should include (1) developing and integrating new sensing technologies and optimizing field deployments across multiple User Facilities to explore interactions and feedbacks across different scales and (2) identifying the emergent processes that serve to couple key subsystems. Such workshops could consider both a “top-down” approach, where User Facility needs are scoped from the Grand Challenges, and a “bottom-up” approach, where the User Facility science priorities are aligned. Specific areas where workshops may be beneficial include (1) biological organization and biosphere-atmosphere feedbacks; (2) human–Earth system interactions (joint with the Energy Sustainability and Resilience theme and potentially feeding into the Network of Energy Sustainability Testbeds (NEST) multifacility management design concept; (3) coupled biogeochemical, energy, and water flows; and (4) atmosphere-land interface.

Recommendation

3.23 Encourage a joint focus on Grand Challenge–relevant scientific themes through (1) coordinated User Facility activities, where linkages are well established, and (2) workshops to develop a vision for coordinated efforts to address cross-disciplinary themes in the Grand Challenges.

Addressing Scaling Issues

Issues of scaling for emergent parameters in both space and time underpin numerous BER Grand Challenges in Earth and Environmental Systems. In addition to the development of additional user facility capabilities outlined on p. 29, there are tremendous opportunities to address scaling through leveraged coordination across User Facilities and activities supported by other agencies. Within BER, as an example, biogeochemical process analyses that might be conducted, primarily via EMSL experiments, could substantially leverage colocated and spatially distributed network measurements from AmeriFlux, ARM, or other facilities to assess the environmental conditions and context that would enable up-scaling. Such coordination might require more directed use of specific facilities or an ability for investigators to effectively access multiple facilities. The upcoming 2019–2020 ARM MOSAiC and COMBLE field experiments offer another unique opportunity to address spatial scaling. ARM will operate mobile facilities within the Arctic sea ice and in northern Norway, in coordination with numerous fixed-station and aerial measurements made by other agencies between these end points to clarify how northward versus southward advection of moisture relative to the Arctic boundary impacts cloud properties and spatial organization.

In addition to scaling in space and time, translating across scales of organization is another critical and ambitious goal: understand how genomics (and other omics), microhabitats, and cellular energetics and thermodynamics interact with and influence processes at ecosystem and Earth system scales. The purview of BER and its User Facilities make them uniquely positioned to address this scaling challenge, which reiterates points raised earlier in this chapter (Bridging Scales, p. 30) and is further developed in Chapter 4: Microbial to Earth Systems Pathways, p. 44.

BER must also consider global scaling through careful interfacing with satellites and models, because many Grand Challenges consider global feedbacks, tipping points, and processes. For example, observations of detailed cloud properties and processes offer the means to quantitatively evaluate satellite-derived cloud products, an assessment which can then serve to provide an informed perspective of clouds globally. This type of interagency coordination requires open sharing of data, the ability to ingest large satellite datasets into BER data systems, and appropriate science funding to support research into scaling. Multiscale modeling is another means for bridging across scales. While projects such as LASSO serve as tools to study Earth system processes and test model parameterizations, stronger links are needed to other tools, from DOE’s E3SM development efforts, to NOAA’s operational forecasts, and to other model center activities. Concerted efforts can be made to examine the potential benefits of assimilation of higher-order parameters from BER networks and observatories (AmeriFlux and ARM) into operational models, and for using these large-scale models to better contextualize individual observational locations. Such efforts would promote coordination of remote-sensing resources and algorithms for synthesizing data important for field activities, model development, and model execution. This could involve a data center for curating fine- to global-scale model results, large observational data from satellite remote sensing, and model-data fusion products.

Recommendation

3.24 Establish stronger links between the operational model and satellite communities to explore more effective transfer of knowledge between BER facilities and these platforms via assimilation, assessment, and intercomparison.

Linking User Facilities and Science

BER User Facilities support and enable cutting-edge scientific research by a broad user community. Moreover, interdisciplinary and inter-facility research activities are being increasingly valued and required by the community. To effectively support current and future BER research priorities and Grand Challenges, User Facilities must explore a new level of coordination that provides a robust means for participation by the science community. Challenges in this process involve alignment among User Facilities themselves, as well as between User Facilities and research funding programs. Additionally, while much of this alignment can be internal to BER programs and facilities, special consideration also is needed for determining how to engage with independent agencies and facilities, as well as other offices within DOE (e.g., ASCR).

Currently, there is disconnect between PI-driven use of BER and other Office of Science User Facilities and the support for related PI-driven science. In some cases, scientific support is gained first, but the scope of the funded project is dependent on the uncertain proposal process for gaining the required access to BER computational resources. For example, projects requiring substantial computational time must write separate proposals for the science (for BER-funded resources) and computing allocation (ASCR facilities) parts of their projects. If the necessary computer allocation is not awarded, then researchers must adjust the scope of the scientific research. On the other extreme, ARM facility deployments do not provide direct support for PI involvement, instead requiring it to be supported through independent proposals to BER's or other agency's science programs. These incongruities in linking facility usage with science support can adversely impact the scope, leadership, oversight, implementation, and scientific impact of User Facility activities. Some consideration should be given to how BER User Facilities can be better coordinated with BER and other science funding programs to enable effective engagement by users and the optimized use of the facilities.

With the trend toward high-impact coupled system research, and in the spirit of addressing numerous interdisciplinary Grand Challenges, there is a tremendous opportunity to directly leverage multiple User Facilities. The barriers to this type of coordination can be significant as a result of distinct timing cycles, implementation considerations, and proposal processes for User Facilities. These barriers may be even larger for coordination with facilities that are external to BER. For example, coordinated access to offshore facilities, such as ships provided by NOAA, may be advantageous for addressing key BER science questions. Additionally, the BER User Facilities can offer a tremendous opportunity for NSF-, NOAA-, or NASA-funded scientists to accomplish cross-agency science objectives. For internal BER activities, the FICUS approach is a successful step in the right direction and can serve as a basis for more advanced coordination involving more User Facilities.

Recommendation

3.25 Further develop and implement a framework for joint calls, review, and decision making (perhaps via the FICUS program): (1) across multiple User Facilities to enable and incentivize cross-disciplinary research to address joint research priorities and Grand Challenges and (2) across User Facilities and appropriate science programs to ensure the availability and effective use of scientific resources. The primary focus for such a framework may be internal to BER, but it should also consider engagement from external agencies and facilities. Such joint calls could be supported through dedicated crosscutting budgets for integrative research.

Engaging the Next Generation

The DOE User Facilities represent an assemblage of state-of-the-science research tools and capabilities identified to deliver the greatest scientific impact to advance the DOE mission. These research tools and capabilities require specific expertise for engagement, while at the same time require evolution and advancement as new ideas and technologies become available. To facilitate the growth of expertise, encouraging new perspectives and technologies, User Facilities must invest in capacity building by engaging the next generation of scientists.

The DOE Office of Science currently manages several programs that encourage growth and engagement of early career scientists, postdoctoral researchers, and graduate and undergraduate students:

- **Early Career Research (ECR) Program**. Supports an annual Funding Opportunity Announcement (FOA) that targets outstanding scientists in the early stages of their career with the goal of stimulating research activities within the DOE Office of Science. These FOAs are generally targeted to specific subdiscipline topics that often are linked to the User Facilities.
- **Office of Science Graduate Student Research (SCGSR) Program.** Provides supplemental awards for graduate students to perform a portion of their graduate thesis research at a DOE laboratory. SCGSR is a direct link for the participating students to the User Facilities and enhances the capacity building that is essential for growth in User Facility science areas.
- **DOE Workforce Development for Teachers and Scientists (WDTS).** Offers internship opportunities for students to develop their mathematical, scientific, and engineering skills by working directly with scientists at DOE's national laboratories. These programs offer invaluable experience for students, often through the User Facilities.

Other important mentorship and training activities are directly guided by the individual User Facilities. For example, BER's TES program funds a postdoctoral researcher at EMSL, and the ARM Facility has previously funded postdoctoral researchers at multiple climate modeling facilities. Another example is the ARM Summer Training and Science Applications event, which invites young scientists for a week of intensive, hands-on education in observations and modeling of aerosols and clouds.

DOE BER can build on these successful educational outreach programs to further strengthen and increase needed capacity and access to new perspectives and technologies. One pathway is to take advantage of recent advanced training and science applications events organized to leverage specific User Facility capabilities, particularly aimed at the Grand Challenge themes. Of particular interest are crosscutting topics that leverage multiple facilities, build interdisciplinary capabilities, and encourage collaborative approaches. User Facilities can also better leverage existing Office of Science educational programs through more direct links between program offices, mentoring activities, and student institutions. Funding for participation of SCGSR and WDTS students in programmatic meetings is one simple way to improve connections between these successful programs and the User Facilities.

Recommendations

3.26 Strengthen the connection of “capacity building” programs with specific User Facilities and specific Grand Challenge themes.

3.27 Consider cross–User Facility summer schools or advanced training activities that bring together diverse groups of students and scientists, organized around leveraging User Facility capabilities for specific Grand Challenge themes.

Chapter 4. Microbial to Earth System Pathways

CHARGE 1 RESPONSE

Alignment of User Facilities to Current Microbial to Earth System Pathways Research

Research supported currently through BER's research portfolio is contributing valuable science directly related to Microbial to Earth System Pathways (MESP). At EMSL, various forms of NMR, electron paramagnetic resonance (EPR), and mass spectrometric (MS) approaches are contributing powerful new insights into the interaction of microbes and soil organic matter (SOM) with minerals, as well as how those interactions affect biogeochemical transformations, proteomics, and the chemical composition of SOM and the small molecules in soil. High-resolution microscopy offers chemical, structural, and cell or organismal imaging at the very small scales important for understanding mechanisms underlying biogeochemical function catalyzed by microbes. Multiscale modeling and high-performance computing are building toward a robust computational framework in which to examine the emergence of larger-scale (e.g., core scale or larger) biogeochemical patterns from the aggregation of pore- and microbial-scale biogeochemical reactions, which are influenced by local conditions, diffusion and advection of resources, and microbial genetic capacity.

At JGI, epigenomics, single-cell genomics, DNA synthesis (synthetic biology), and metabolomics are powerful new additions to already-established basic sequencing approaches (e.g., metagenomics, metatranscriptomics, and meta-barcoding). These techniques lay the foundation for delving deeply into the genetic underpinnings of microbial function and its dynamic control, including genotype-phenotype relationships and how functionality responds to acute and chronic perturbations. The informatics and data science required to analyze enormous datasets from these experimental approaches have contributed to driving a strong partnership with NERSC for high-performance computation, and this partnership will continue to grow in importance as high-throughput analyses enable exploration of heterogeneity in biological process in populations of cells and communities of organisms. That biological heterogeneity contributes to multifaceted phenotypic responses that collectively support the capacity for biogeochemical process underlying resiliency in ecosystems and the possibilities for thoughtful harnessing of microbial capabilities in combating environmental stressors.

JGI, EMSL, NERSC, and the advancements in KBase all are providing foundational contributions to multimodal measurement and modeling that are building toward better understanding and prediction of Earth system function across scales. The Center for Integrated Nanotechnologies offers the ability to construct model environments in which to investigate targeted microbial activities. Complementary approaches for characterizing complex, heterogeneous materials, and processes occurring within them (e.g., carbon stabilization in soil), are provided by diverse X-ray and infrared imaging and crystallography at the Advanced Light Source (ALS). At larger scales in the field, AmeriFlux measurements offer direct links between biosphere and atmospheric function, complementing JGI's and EMSL's joint biological, hydrological, and soil or sediment expertise. ARM's mobile facility contributes as well, such as concerning site-based soil moisture and precipitation and through targeted campaigns as in the Southern Great Plains. DOE's targeted campaigns (e.g., NGEE-Arctic, NGEE-Tropics, and SPRUCE) provide even more detailed, biosphere-focused, site-specific measurements, and associated modeling approaches; these campaigns provide important field context for honing conceptual advances based on mechanistic understanding developed under more controlled settings.

There are, however, impediments to establishment and advancement of User Facility projects supporting Grand Challenge–focused research.

Greater Integration of Experimentation and Modeling

At times there are obstacles to the integration of users into the development of cutting-edge process modeling, informed by experimentation, at the frontiers of microbial habitat-scale science. Yet that integration of experimentation and modeling, as well as the cross-fertilization of ideas between national laboratory EMSL personnel and users, is critical for advancement toward the Grand Challenge of predictive understanding across scales. Partnerships among experimentalists, computer scientists, and modelers are required and an institutional framework that reliably supports development of such partnerships needs to be in place.

As an example, through FICUS and EMSL Science Area calls, EMSL offers access to the CASCADE supercomputer. User projects have been funded to use established process models, such as STOMP, e-STOMP, and NWChem, and to take advantage of the ongoing development of the pore-scale modeling framework on CASCADE. Allocation of CASCADE time is a fundamental part of the proposal approval process. Historically, EMSL computing has emphasized computational chemistry, and the expertise of user support staff is mostly in that area. EMSL also has recently developed strong user support for analysis of metabolomic and other omic data. However, user projects led by experimentalists striving to link with models in BER science areas other than computational chemistry or bioinformatics can languish for lack of modeling expertise. EMSL can draw on Pacific Northwest National Laboratory (PNNL) staff modeling expertise, but doing this often proves challenging because of other priorities of non-EMSL staff and can hinder effectiveness of university–national laboratory collaborative projects; it also is in contrast to EMSL's excellent tradition of providing expert technician support for projects using, for example, microscopes, MS, and NMR, and in contrast to the effort EMSL has been making to provide increased support for data analysis and interpretation.

EMSL expansion to include the array of applications and codes relevant to BER users is essential because effective integration of computation and experiment goes far beyond simply providing access to high-end computing hardware (e.g., CASCADE). Expansion to tackle focal projects could be accomplished in part through a cluster of postdoctoral fellowships; mentorship and user collaboration through an expanded support staff also will be critical. As is already required by EMSL, user proposals including CASCADE-based process modeling time should be thoroughly discussed with appropriate computing personnel prior to submission. If the work is chosen, a clear, standard mechanism for users to obtain the computing and/or modeling guidance (that was agreed on prior to submission of the proposal) should be in place. (This is particularly important for model and code development that is ongoing, with shifting personnel involvement.) Such collaborations during the development of new code (e.g., multiscale approaches) encourage adoption of cutting-edge computing approaches by a broad research community rather than being predominantly confined to users that already have a high level of computational expertise. Entraining new users will bring novel perspectives to the modeling that computer scientists otherwise might not access through their national laboratory colleagues.

Recommendation

4.1 Expand EMSL computational support staff and their expertise to include the array of applications and codes relevant to BER users.

Data Release Rules

Immediate data release rules at some User Facilities can encourage minimal metadata reporting and can deter establishment of inter-Facility, interdisciplinary projects when data production and release are on different timetables at different Facilities.

The immediate release of data by some DOE User Facilities puts users who have conducted experiments and provided samples at a disadvantage if they have mixed teaching-research appointments at universities and colleges without large computational support groups, or if they are early career researchers just becoming established. These users potentially find themselves competing to publish their data before the information is downloaded and enveloped by one of many well-funded computation groups at other institutions. Though the User Facilities indicate this should not happen, it is enough of a threat that it sets up a dynamic whereby users may provide the absolute minimum of metadata to deter use of the dataset before users can publish it. This diminishes the value of the dataset in the long term for the broad research community. Immediate data release is often justified as an effective way to stimulate the pace of science, but the reality is that the quality of science can suffer without the benefit of leadership by those scientists who best know the organisms and the conditions and questions that drove the careful experimental design and collection of the data. The effect of immediately releasing the data is to channelize analysis and interpretation into particularly well-funded groups with access to the analytical and computational tools to rapidly convert new data to published papers. Not only does this negatively affect the careers of scientists with less access to those resources, the papers that emerge can be detrimental to the science, lacking the benefit of more informed, careful scientific interpretation.

An important goal in support of Grand Challenge research is the melding of information from multiple perspectives and scales. Programs such as FICUS support inter–User Facility projects. For single awards, however, the length of time awarded for single projects being conducted at more than one User Facility can differ at each facility. Data release rules also can be different. These mismatches put users in the difficult position of defending data released by one Facility prior to availability of data from the other, or defending a delay in production of paper(s) after an award has ended at one Facility but not the other.

Recommendation

4.2 Institute a time delay before data are released, until publication, or for one year after a user project ends, whichever comes first. For projects with components at different User Facilities, match the time frames of the project components as well as the time delays for data release.

Sample Analysis Throughput

Sample analysis throughput has not been able to keep up with user demand, with a significant lag in data delivery. Also, the data themselves have become more complex, leading to increased need for computational assistance from the User Facilities generating the data.

As User Facilities take on new scientific challenges that bring unprecedented volumes of data, the need for advanced computing to support users' scientific work after measurement is increasing rapidly. User Facilities are barely meeting the demands of the BER-relevant user community in current Microbial to Earth System Pathways research—sample and data processing times are not as quick as some users need or would like. For example, introducing postdocs to the User Facility offerings is in the interest of User Facilities, but postdocs are often on a very tight 2-year schedule for sampling, obtaining data, and composing a report to support next steps in their careers.

User Facility success encompasses maintaining a productive, returning user base. Metrics should take into consideration this measure of success rather than weighting toward total numbers of users served. Implicit in the ability to keep users engaged is the ability of the User Facility to keep turnaround times for project completion rapid enough to match user needs at different career stages. If metrics of User Facility success are too strongly influenced by numbers of users, pressure will develop for Facilities to increase user membership to the detriment of speed of project throughput and/or depth of collaborative analysis. Clearly, different kinds of analyses require different amounts of time, so no one size fits all policy for data delivery.

User Facilities writ large are poised to take leading roles in developing an information technology emphasis contributing to multiple aspects of BER Grand Challenge research. Early work of KBase, JGI's Integrated Microbial Genomes and Microbiomes (IMG/M) and EMSL software to identify chemical compounds reported by mass spectrometry are all examples obviously relevant to linking Microbial to Earth System Pathways. Across the broad range of BER interests, the following capabilities are critical: data curation (storage, safety, formatting, and plotting), broadly understandable user interfaces (retrieval and transfer), and downstream computing converting primary data into value-added data. These activities will require computing resources and personnel, and they clearly are opportunities for User Facility leadership in the broad scientific community.

Recommendation

4.3 Shift weight toward metrics of User Facility success that recognize facility efforts in maintaining a productive, returning user base, rather than weighting toward total numbers of users served.

CHARGE 2 RESPONSE

Alignment of User Facilities to Address Future Needs and Grand Challenges in Microbial to Earth System Pathways

We recommend that User Facilities employ specific strategies to address the four Grand Challenges identified for understanding links from Microbial to Earth System Pathways:

Grand Challenge 4.1: Characterize the biogeochemical exchanges driven by food web and plant-microbe interactions and evaluate their process-level impacts, sensitivity to disturbances, and shifting resource availability under changing environmental regimes.

Grand Challenge 4.2: Define the sphere of influence and key elements of microbial communities in space and time relevant for predicting larger-scale ecosystem phenomena for Earth system understanding.

Grand Challenge: 4.3: Integrate molecular and process data to improve the ability to define ecologically significant traits of individual taxa and communities and use trait-based models to develop predictive links between community dynamics and ecosystem processes.

Grand Challenge 4.4: Align and deepen connections among conceptual understanding, measurements, and models related to the roles of microbes in determining the rate of transformation, uptake, and loss of chemical elements from ecosystems.

As a result of their collaboration with diverse users, User Facilities are in a particularly strong position to contribute to the development of scaling rules (perhaps even scaling laws) and computational strategies that can link microbial processes and their consequences across scales of time, space, and complexity. Discerning scaling principles is essential for grappling with the nested systems, nonlinearities, feedbacks, and coupled processes that are common in Earth systems. For this effort to be successful, User Facilities must be equipped with instrumentation and computation facilities as well as sufficient personnel time to develop the necessary framework in collaboration with users who bring diverse expertise. (See section titled "Development of Additional User Facility Capabilities," p. 50) Mechanisms could include cross-Facility research calls emphasizing interdisciplinary collaboration and/or exploration of linkages across scales, postdoc cluster hires focused on particular problems, and collaborative short-term (jamboree-like) or longer-term (12- to 18-month) user–User Facility working group collaborations.

The recommendations for addressing the Grand Challenges include exploring diverse scaling strategies, developing a multiscale framework for interactive modeling and experimentation, and fostering interdisciplinary interactions.

Explore Diverse Scaling Strategies

A focused effort is required to develop mathematical scaling rules and related principles integrating observations across spatial scales (e.g., nano-, micro-, and mesoscale) and dynamics or process rates across timescales.

Scaling is a well-recognized necessity in ecosystem research, but efforts toward implementation of ideas in the existing scaling literature, and new ideas emerging from multiscale modeling efforts, have been hindered by the difficulty of the charge. For example, Chapter 4 in the 2017 BERAC Grand Challenges[13] document emphasizes the promise of trait-based approaches for distilling the essence of microbial function to support scaling (e.g., Grand Challenge 4.3). However, trait-based modeling dominantly focuses on the "bio" and microbial genetic capacity, which are critical components but are not sufficient to predict biogeochemistry. The local environmental ("geo") conditions and resources also influence reactions and rates, but unclear is how microenvironmental and microbial community heterogeneity at tens of micron scales (i.e., the microbial "habitat" scale) influence spatial and temporal patterns in processes at far larger scales.

A diversity of approaches to scaling, supplied by a variety of users, should be enlisted to contribute to this scaling effort, hand in hand with experimental and observational data. For example, scaling based on thermodynamic opportunities and constraints is gaining momentum, especially when combined with information based on known microbial capabilities. Another fruitful approach may be to explore scaling of Microbial to Earth System Pathways within the framework of complex systems theory; in that context, establishment of scaling laws requires conceptual development to deal with open systems, nonlinearities, nested systems, feedbacks, networks, coupled processes, and emergent properties. A variety of computational scaling testbeds, coupled iteratively with experimental tests, will be most fruitful moving the field forward.

Develop a Multiscale Framework for Interactive Modeling and Experimentation

A research thrust focused on scaling necessitates development of a robust framework that is flexible enough to accommodate experimental data input and integrate process modeling across scales.

This computational framework should serve as a scaffold enabling focused collaboration within groups targeting particularly challenging problems, as well as enabling experimental and observational activities to be exploited by modelers, and vice versa (Grand Challenge 4.4). Multiscale understanding will require building and linking interdisciplinary research communities and equipping them with tools for iterative integration of measurements and modeling.

Foster Interdisciplinary Interactions

Building a robust framework linking microbial functions to Earth system pathways necessitates strong cross-pollination among seemingly distant disciplines; User Facilities are an ideal environment to attempt combined activities of this kind.

As emphasized above, biogeochemistry catalyzed by microbes is a function of both microbial genetic capacity and environmental opportunity. Significant parts of the subsurface biosphere can be seen as a natural bioreactor—a porous environment inhabited by organisms that process substrates at variable rates. To completely describe and understand the workings of such systems, information about organismic (i.e., genetic) capacity and its expression must be accompanied by information about reactor design (i.e., habitat properties). To use this knowledge for predictive modeling, collaboration among molecular biologists, soil physicists (e.g., new soft condensed matter physicists), hydrologists, ecological theoreticians, computer scientists, and others will be essential. At present, capabilities at User Facilities have matured with foci on certain scales and are poised to link with other agencies and groups with complementary specialties. For instance, EMSL has hardware and expertise for investigations at the spatial

[13] BERAC. 2017. *Grand Challenges for Biological and Environmental Research: Progress and Future Vision; A Report from the Biological and Environmental Research Advisory Committee*, DOE/SC–0190, BERAC Subcommittee on Grand Research Challenges for Biological and Environmental Research. (science.energy.gov/~/media/ber/berac/pdf/Reports/BERAC-2017-Grand-Challenges-Report.pdf)

scale of the microbial habitat, ARM collects data at both the particle (aerosol) scale as well as the regional scale, and EMSL and JGI have expertise in handling highly complex microbial and plant omic data. Facilitating information flow among disciplines and across User Facilities is a very promising avenue for the future. Suggestions for ways to engage partners in the effort are made in the sections that follow. Because productive interdisciplinary collaborations are challenging to establish, we also suggest ways to establish and grow the user base trained to think broadly and link across fields.

CHARGE 3 RESPONSE

Development of Additional User Facility Capabilities

Enable Process Modeling and Data-Related Computation

New investments are needed in midrange computing infrastructure and in personnel time to enable process modeling and data-related computation. Achieving the Grand Challenge goals for linking Microbial to Earth System Pathways will require that experimentalist users with extremely diverse scientific backgrounds collaborate with computer scientists and modelers. There is precedent for such a targeted collaboration; for example, PNNL/EMSL shepherded the development of NWChem[14] in response to user and community need for a computational framework to analyze kinetics and dynamics of chemical transformations generally and, more specifically, the chemistry at interfaces[15] and in the condensed phase below ground. This chemistry-oriented community had more extensive computational background; the BER users likely to contribute to the development of the experiment- and observation-informed multiscale modeling framework envisioned here will be less grounded in computer science and process modeling. Therefore, personnel time for computer scientists and process modelers must be allocated to this effort, not just CASCADE time or access to developed equations. Provision of compute time with associated personnel time is essential for fostering the iterative, interdisciplinary vision that is required to reach Grand Challenge goals.

Similarly, experimentalists collaborating with EMSL scientists often do not have the training to analyze very large datasets on their own. We make some recommendations for training current and next-generation users in the Charge 4 section of this report. From the perspective of experimentalists receiving increasing volumes of more and more complex data, provision of assistance with data management, analytics, visualization, and interpretation have become essential parts of the user–User Facility relationship. JGI's growing partnership with NERSC is very promising for large-scale data analysis, modeling, and storage. New investment in midrange computing infrastructure (integrated central processing and graphics processing units [CPUs and GPUs], hardware, software, and personnel expertise) at EMSL is required to support future user–User Facility partnerships focused on integrated "bio" and "geo" controls over microbial catalysis of biogeochemical processes across scales.

Computational Network for Connecting and Informing Multiscale Models

The magnitude of the Grand Challenge to link from Microbial to Earth System Pathways requires development of a large collaborative body of user and User Facility researchers working toward a common goal. Development of a robust computational framework to support this collaborative body is a very complex task and requires support for focused personnel time as well as equipment and computing time.

---

[14] NWChem – High-Performance Computational Chemistry Software (www.nwchem-sw.org/index.php/Main_Page); accessed October 2018.

[15] U.S. DOE. 2017. Research Priorities to Incorporate Terrestrial-Aquatic Interfaces in Earth System Models, DOE/SC-0187, U.S. Department of Energy Office of Science (tes.science.energy.gov/files/TAI_Workshop2016.pdf).

Establishment of the framework will at best be slowed and at worst fail if it is piecemeal. Given that the BER mission is one of system science spanning spatial and temporal scales with the explicit goal of predictive understanding, and given the more targeted goal of linking Microbial to Earth System Pathways, closing the gap is essential to understanding at the microbial habitat scale. The computational challenge and opportunity in BER space is linking experiment and computation (both simulation modeling and data analytics). The goal is integration capability, not a program that overlaps or competes with computing powerhouses NERSC and Oak Ridge National Laboratory (ORNL). Rather, hand-in-hand empirical and computational effort will be the foundation and unusual strength of this collaborative framework.

Framework development and framework access will need to be open to the broad user community, and we would encourage that a variety of approaches to scaling and integration be encouraged and tested. Creative breakthroughs may emerge from embracing diverse ideas for scaling from Microbial to Earth System scales, whether those approaches are genomic, thermodynamic, “trait”-based, completely novel mathematical, or some hybrid of all of the above. An important component will be to identify any emergent properties and scaling “laws” resulting from the work.

As one example, at the microbial habitat to pore, core, and plot scales, EMSL can provide the platform for framework development and integration. The computing architecture necessary to develop and implement this dynamic, collaborative framework would be dictated by current user and framework development needs. Those needs currently call for both midrange CPU and GPU hardware for process modeling and simulations, as well as for data analysis and management. As community tools are developed by community members, those tools can be linked into the growing modeling framework. Standard models (developed at different scales) could be arrayed within the framework to pass information back and forth, not solely upward from the most mechanistic up to the largest-scale model. Patterns emerging from experiment results and modeling output at multiple scales would suggest scaling principles and rules that could be further tested. Among the pieces of the framework that have already become established could be, for example, (1) ELM (Export Land Model linking plot to globe via E3SM), (2) MEND (a microbially explicit carbon model, potentially linking core to plot), (3) PFLOTRAN (a macroscopic flow and reactive transport model linking core to plot scales), (4) TETHYS (a pore-scale flow and reactive transport model), (5) KBase (capturing cell and microbial community metabolism), and (6) NWChem (operating at the molecular scale). The ultimate goal is a modeling framework linking across scales, with classes of models plugged into a standardized interface facilitating interoperability and cross-scale information flow.

The result will be something like a “virtual laboratory[16]”—a facility that integrates and iterates between lab and field experimentation (at multiple scales) and process modeling and data analytics (again, at multiple scales). Community buy-in will be critical. These integrated activities should not take place in a new facility that is separate or isolated from existing User Facilities; the integration, interdisciplinarity, and iteration of experimentation and modeling are critical points to the effort. “The essential three” are best achieved in physical or virtual units where scientists of both ilk interact, creating a focused support structure for scaling biogeochemical processes. Input from multiple User Facilities would be essential to scale from Microbial

The Essential Three:

Integration

Interdisciplinarity

Iteration

[16] 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/~/media/ber/berac/pdf/Reports/BER_VirtualLaboratory_finalwebLR.pdf)

to Earth System processes. As outlined in the example above, EMSL can provide a framework spanning microbial habitat to core, and perhaps even to plot, scales. JGI (and KBase) can provide critical empirical omic data that will help discern which genomic and community details are and are not essential for modeling biogeochemical processes. Links will then need to be built to other larger-scale empirical measures and models (e.g., to scientists in ARM, AmeriFlux, and NEON), as well as to ELM and E3SM[17] in the Climate and Environmental Sciences Division[18] (CESD) of BER, and to CESM in the National Center for Atmospheric Research (NCAR); managed by the University Corporation for Atmospheric Research (UCAR) on behalf of NSF's UCAR Climate and Global Dynamics Group.[19] Goals are already complementary in these various groups; they are poised for synergy. For example, for E3SM, the goal for the Biogeochemistry focus is to address the fundamental question: "How do biogeochemical cycles interact with global climate change?" E3SM simulates the Earth system at 15- to 25-km resolution, coupling carbon and nutrient cycles (and feedbacks between them), as well as energy, water, and land use for hind- and forecasts.

Accomplishing this integration requires that links be made across scales, modeling platforms, and user and User Facility communities. Clear line-of-sight plans must be developed for tractably incorporating experimental and modeling work from diverse communities in an iterative process, and for passing information up and down through models poised at various scales for testing, for example:

- What complexity at small scales matters at large scales, and what complexity does not?
- Which major uncertainties in dynamics, thresholds, and feedbacks need to be better understood and captured in models in order to link across scales?

Finally, for this framework development to flourish now and with input from the next generation, users need to be conversant in both empirical studies and modeling and computational work. There are opportunities for interdisciplinary and interactive training that supports both framework development now and invests in personnel development for the future.

Field Deployable, Multimodal, Remotely Controlled Sensors

Combining lab-based, mechanistic measurements with multiscale modeling will provide a window into field mechanisms potentially important for linking Microbial to Earth System Pathways. Those hypothesized field mechanisms need to be tested, and real-time measurements of field processes at multiple field scales will be critical to that effort. The development and deployment of field-worthy, multimodal, remotely controlled sensors are thus essential for ensuring that lab-based experimentation and multiscale modeling *in silico* can progress iteratively with field insights to produce the most useful representations of links from Microbial to Earth System Pathways. Particularly for elucidating fundamental mechanisms influencing Microbial to Earth System Pathways, there is a notable gap in our understanding of how heterogeneity and dynamics in microbial communities, and in the resources and conditions in the local microbial habitat, influence larger-scale processes in the field. Installations might be semipermanent and/or mobile. In ocean systems, for example, field-deployed, automated samplers are now describing microbial communities through time. Do the realities of operation in soil necessarily preclude a similar

---

[17] Energy Exascale Earth System Model: climatemodeling.science.energy.gov/projects/energy-exascale-earth-system-model/; accessed October 2018.

[18] U.S. DOE. 2018*. Climate and Environmental Sciences Division Strategic Plan 2018–2023*, DOE/SC-0192, U.S. Department of Energy Office of Science (science.energy.gov/~/media/ber/pdf/workshop%20reports/ 2018_CESD_Strategic_Plan.pdf).

[19] NCAR UCAR Climate & Global Dynamics; 2015 Annual Report: nar.ucar.edu/2015/cgd/multiscale-modeling-systems/; accessed October 2018.

approach? Could a first step, for example, be sampling of the mobile microbial community in soil solution, through time, since this mobile community advected through soil can be the inoculum for downstream ecosystems? Furthermore, discovery of biogeochemical "hot spots" and "hot moments" in the field, and their balance in space and time with "cold" spots and moments, should also help with identification of process triggers and tipping points important for model development to inform larger scales. Can new, nondestructive, *in situ* observation systems (e.g., image based) broaden continuous measurements beyond point samples to at least 2D dynamics through time? For broad adoption by the scientific community (potentially, also by "K-grey" educational partners for public outreach), ideally a subset of deployed sensors would also be low cost, low energy, low maintenance, and capable of "smart" function (e.g., triggered by target events).

Opportunities to link with existing ecosystem-scale experimental campaigns and continental- and larger-scale networks of field instrumentation (e.g., AmeriFlux and NEON) are described earlier in this chapter." User Facility science calls should make clear that coordinating with these networks is within the purview of user projects. If establishing a complementary network of next-generation sensors is a long-term goal, learning from the experiences (positive and negative) of established networks is critical. Organization of such field science networks requires clear leadership and establishment of standards to facilitate comparisons across sites. For example, base measurement expectations must be established, but also there must be cognizance of site-specific peculiarities, and thus flexibility based on expert input and units; organization of data reporting should be standardized, and there should be agreement on quality assessment/quality control (QA/QC) standards).

Recommendations

4.4 Enable process modeling and data-related computation by investing in midrange computing infrastructure and personnel time.

4.5 Develop a robust computational framework that can connect and inform models at multiple scales and that facilitates iteration based on input from experimental and field data and modeling output.

4.6 Develop field deployable, multimodal, remotely controlled sensors that ideally conduct nondestructive measurements to (1) characterize how microbial habitat–scale heterogeneity and dynamics influence biogeochemical processes and (2) validate relevance of lab experiments in field.

CHARGE 4 RESPONSE

Opportunities for Collaboration Among User Facilities

Collaboration among User Facilities, groups within DOE, and interagency partners is not only a very real opportunity but is in fact a necessity to develop a robust framework for understanding links from Microbial to Earth System processes.

Contributions from EMSL, JGI, NERSC, KBase, ARM, NSF's UCAR NCAR, NGEE-Arctic, NGEE-Tropics, and SPRUCE are all noted in sections above. Collaborative partnerships with long-term ecological networks both within and outside of DOE, where extensive suites of process measurements are planned for decades (e.g., NSF's NEON and LTER and DOE's AmeriFlux), will provide data essential for describing the temporal and spatial variability of environments, and the processes influencing, and influenced by, microbial systems. For example, at the continental scale and to a lesser extent around the globe, we will soon have near-continuous measurements of CO_2 and water fluxes from a large number of ecosystems, with data on $^{13}CO_2$ and CH_4 from a smaller number. There are also quite good satellite measurements of plant parameters (NDVI), as well as targeted intensive airborne measurements of trace gases and hyperspectral data. In addition to observational and monitoring data from these entities, there are numerous manipulative networks at the ecosystems scale that could contribute field measurements as testbeds for multiscale modeling and grounding of lab-based mechanistic studies. DOE's national laboratory User Facilities have the opportunity to leverage these extensive ecosystem-scale datasets, sample archives, and simulations to both inform and integrate the microbial habitat scale into biogeochemical understanding. The existing ecosystem network provides a rich resource for generating new hypotheses and developing models that can contribute to the multiscale collaborative framework.

Ensuring improved communication and coordination among facilities and networks will be essential for recognizing and capitalizing efficiently on opportunities for synergy. Calls for User Facility proposals may need to be worded to more explicitly reach out to communities where synergies likely lie but community members are unaware of User Facility opportunities.

User Facilities can offer tremendous opportunities for interdisciplinary, interactive training that supports both the development of a multiscale, collaborative framework now, and that invests in critical personnel development for the future.

User Facilities and their activities have the potential to nucleate data analytic, data synthesis, and model development activities relevant to Grand Challenge science via a number of mechanisms, such as the following.

Synthesis Workshops and Campaigns

Synthesis workshops bringing together existing Facility users and outside scientists could be a catalyst for the creative thinking that will accelerate multiscale model-experiment innovation. These workshops could take a variety of forms and be modeled after the intensive 2-week JGI-style "jamborees," or the longer 12- to 18-month collaborations among the National Center for Ecological Analysis and Synthesis (NCEAS), National Socio-Environmental Synthesis Center (SESYNC), and Powell Center dedicated to in-depth analysis of complex problems. There is precedent already for this kind of activity; for example, from 2012 to 2015, a very productive on-site "research campaign" for development and application of pore-scale modeling was conducted at EMSL.

Training Interdisciplinary Scientists

User Facilities also have enormous potential for training scientists interested in reaching across disciplinary lines and for entraining the next generation of environment-oriented empiricists and modelers. These new researchers will become tomorrow's user base for the User Facilities. They will have valuable ideas about which types of research they hope to develop, as User Facilities set strategic goals for the future.

Short Courses at User Facilities and/or National Meetings

User Facilities can contribute to building a common language across disciplines via short-course training. Mixed-discipline attendance will support participants becoming conversant in each other's fields as well as uniting them as a cohort learning a particular empirical or computational skillset. For example, EMSL, JGI, and/or ARM could set up a *2-week crash course* (e.g., like the University of Utah Stable IsoCamp Course[20]) on a topic of their choice each year, to run on their own campus. Furthermore, they could run shorter

[20] Stable Isotope Biogeochemistry and Ecology: stableisotopes.utah.edu/about.html; accessed October 2018.

workshops at national meetings, such as the Ecological Society of America (ESA), the Association for the Sciences of Limnology and Oceanography (ASLO), or the American Geophysical Union (AGU) meetings. Running short courses at User Facilities sites, or offering them at national meetings, would increase DOE's visibility to potential users who otherwise may be deterred by the challenge of even understanding what is offered at User Facilities, let alone getting up to speed in the diverse empirical and computational approaches in use. Topics could rotate (to spread the notable teaching load among User Facilities) or stay the same, session to session (to make the preparation easier, through repetition). Examples might include sessions exploring the opportunities and data analytic and/or process modeling challenges in the following measurement processes:

- NMR and EPR (organic matter, metabolomics, and metabolic flux)
- Microscopy (transmission electron [TEM]; Cryo-EM and Dynamic TEM; electron microprobe; and/or confocal, super-resolution, fluorescence lifetime, stochastic optical reconstruction microscopy [STORM], and photoactivated localization microscopy [PALM])
- Mass spectrometry (laser ablation, Fourier transform ion cyclotron resonance [FTICR], and aerosol)
- Biogeochemical modeling (contributing to the multiscale collaborative framework, introduction to ELM or E3SM)
- Proteomics and Metabolomics (What's possible, what isn't, and why?)
- Synthetic Biology (from idea to implementation)

DOE Postdoctoral Fellowships

DOE needs to build a workforce and user base trained in this new multiscale genre of science. Clusters of postdoctoral fellowships could be offered within or across User Facilities (1) to recruit diverse, talented fellows as a team to work on complex, interdisciplinary problems or (2) to support individual postdoctoral projects that require cross-training across Facilities, thus ensuring broad exposure to diverse ideas.

Chapter 5. Energy Sustainability and Resilience

Energy production and use are inherently connected to land, air, and water resources. Comprehensively understanding these interactions is therefore important for guiding current and future energy production, conversion, and use systems in ways that will appropriately balance energy availability and cost, resilience and security, and environmental quality. This in turn will ensure the sustainability of future energy systems, including their effects on air, water, and land resources. This general concept of "energy sustainability" was one of the major research areas covered in the 2010[21] and 2017 BERAC Grand Challenges reports.[22] To highlight the highest-priority research infrastructure needs of this research community, this chapter moves beyond the broad concept of "energy sustainability" to identify an urgent need for research focused more specifically on energy and environmental resilience. This sharpened focus ultimately leads to a strong case for the recommendation of establishing a network of geographically distributed Sustainable Energy and Environmental Research Centers, akin to BER's current Bioenergy Research Centers (BRCs) and eventually a Coordination, Integration, and Visualization (CIV) Center, which would operate like existing BER User Facilities. Establishment of these new institutions would greatly enhance BER's ability to address the Grand Challenges identified in these past reports, would leverage existing User Facility capacities, and would facilitate new collaborations. For these reasons, discussions in this chapter blend the three components of the charge letter: alignment with the Grand Challenges, new capability development, and new collaborative opportunities (see Appendix A, p. 84).

The nation's energy system is increasingly interconnected with other human and environmental systems. It also is increasingly exposed to a variety of acute shocks (e.g., droughts, floods, heat waves, and wildfires) along with persistent, longer-term changes in energy, water, and agricultural demands resulting from rising populations and incomes, technological changes (e.g., fracking and better batteries), and aging infrastructure. Together, these shocks, interdependencies, and external drivers create risks for the energy system, as well as the potential for cascading failures across water, land, and other interconnected infrastructure systems. Local and regional planners, like those at electric and water utilities, and metropolitan resource management agencies are increasingly being challenged to prepare for and take steps to reduce these risks. Yet, we do not fully understand the complex system dynamics that underlie infrastructure fragility at and across local and regional scales. Moreover, we do not fully understand how these risks may be reduced or exacerbated by different approaches for managing these interconnected systems. Thus, there is an urgent need to better understand how the complex, multiscale dynamics associated with multiple stressors, cross-sectoral interactions, and management approaches could fundamentally alter the vulnerability, reliability, and resilience of energy-water-land systems at both the urban and regional scales. Additionally there are significant interdependencies among energy and environmental systems and changing and/or at-risk natural resources, ranging from water and ecosystem services to forest products. Groundwater depletion and large-scale landscape disturbances of forest and cellulosic stores brought about by insects, such as bark beetle, and infectious diseases of plants are on the national radar, possessing many research components within BER's wheelhouse, which spans both the

---

[21] BERAC. 2010. *Grand Challenges for Biological and Environmental Research: A Long-Term Vision; A Report from the Biological and Environmental Research Advisory Committee March 2010 Workshop*, DOE/SC-0135, BERAC Steering Committee on Grand Research Challenges for Biological and Environmental Research. DOI: 10.2172/1006492.

[22] BERAC. 2017. *Grand Challenges for Biological and Environmental Research: Progress and Future Vision; A Report from the Biological and Environmental Research Advisory Committee*, DOE/SC–0190, BERAC Subcommittee on Grand Research Challenges for Biological and Environmental Research (science.energy.gov/~/media/ber/berac/pdf/Reports/BERAC-2017-Grand-Challenges-Report.pdf).

Biological Systems Science Division and Climate and Environmental Sciences Division, and are critically important for energy and environmental security.

In assessing the value of this type of resilience-oriented energy and environmental research, it is important to recognize the very strong linkages between the questions it seeks to answer and the challenges increasingly being faced by those in other parts of the national research establishment. This would include challenges faced not only by agencies responsible for water, air, and land resource management and infrastructure, but also, importantly, those responsible for both homeland and international security. Perspectives and resources (especially data and relevant research) from these agencies should be major inputs into the development of BER research and research infrastructure in this area.

CHARGE 1 RESPONSE

Alignment of User Facilities to Current Energy Sustainability and Resilience Research

DOE currently invests in a number of activities that directly or indirectly address energy and environmental resilience, broadly considered. For example, there are major programs in DOE's Office of Electricity focusing on the reliability and resilience of the nation's grid. Similarly, there is a range of programs across the Office of Energy Efficiency and Renewable Energy, Office of Fossil Energy, Office of Fusion Energy Sciences, and others that strive to improve efficiency and/or develop and deploy cleaner sources of energy to reduce environmental impacts. Within BER, ARM supports improvements for understanding atmospheric processes, land-atmosphere interactions, and ultimately the development of comprehensive Earth System Models (ESMs), such as the Energy Exascale Earth System Model (E3SM). These tools are critical for projecting the potential vulnerability of future energy systems to both short-term shocks and long-term changes in the Earth system, as well as the impact of energy systems on surface, subsurface, and atmospheric systems. Some of the research at EMSL and JGI also supports energy and environmental resilience, for example, through the development of advanced biofuels and other technologies that potentially could increase options for a more sustainable and resilient energy system.

A key gap that emerges is a research infrastructure that can account for the many interdependencies of individual energy technologies with each other as well as with other human and natural systems. In particular, there is a need to better understand the potential resilience of different energy strategies across a range of future scenarios, as well as the impact of those strategies on other human and environmental systems. This understanding will require major advances in understanding and simulating interdependencies, nonlinear behaviors, and risk modalities across a huge range of systems, as well as data gathering, model development, and analytical efforts across a wide range of sectors and spatial scales. It also is strongly aligned with DOE's mission to ensure the security and prosperity of America by addressing its energy and environmental challenges through transformative science and technology solutions.

DOE has already established an effective model for attacking large-scale research problems on this scale, namely, the establishment of major networks of "research centers" addressing different aspects of the challenge but with a common integrating goal. BER's BRCs, for example, were established to "develop the science, technology, and knowledge base necessary to enable sustainable, cost-effective production of advanced biofuels and bioproducts from nonfood plant biomass in support of a new biobased economy.[23]"

[23] DOE Bioenergy Research Centers: https://genomicscience.energy.gov/centers/BRCs_2018LR.pdf.

Each of the four BRCs focuses on different aspects of this challenge with interdisciplinary teams and scientific strategies that reflect both the scientific diversity and regional heterogeneity of biofuel development and production. Similarly, the Energy Frontiers Research Centers (EFRCs) established by DOE's Basic Energy Sciences Program, were created because "a new fundamental understanding of how nature works is necessary...in order to meet the global need for abundant, clean, and economical energy.[24]" Each EFRC is a partnership among universities; national laboratories; nonprofit organizations; and for-profit firms using a powerful new generation of tools for penetrating, understanding, and manipulating matter on atomic and molecular scales. Any serious effort to address the nation's energy and environmental resilience will demand a similar level of effort and strategic deployment of resources.

CHARGE 2, 3, AND 4 RESPONSES

Alignment of User Facilities to Address Future Needs and Grand Challenges in Energy Sustainability and Resilience: New Capacities and Collaborations

In recent years, research on bioenergy conversion and associated environmental considerations has progressed substantially, accompanied by an increased understanding of energy-food-environment tradeoffs and improved characterization of spatial and temporal variabilities of targeted ecosystems. In addition, progress has continued on understanding the linkages between fossil-fueled and nonbiofueled renewable energy systems and water, air, and land systems. Significant advances have included further development of multisector dynamic models, climate models, integrated ESMs, impact and vulnerability models, and the coupling of these models where appropriate to fully address sustainability science questions. Moving forward, four Grand Challenges will take this research into the next decade and help resolve important questions:

- Further develop the science of coupling energy, water, and land use across different spatial and temporal scales to understand environmental impacts and changing climate and to better predict net biogeochemical fluxes.
- Use observational, experimental, and model-based approaches to explore the sustainability of alternative energy systems, incorporating stability and resilience analysis, uncertainty, transition paths from current infrastructures, and the use of appropriate common metrics.
- Understand how variability and change in natural systems affect energy system structure and function and determine how best to build this knowledge into models.
- Create new data streams and use existing observations more effectively to ensure the availability of scale-appropriate data, particularly related to high-resolution land use, landscape infrastructure, demographic change, and energy-land-water research.

The Grand Challenges report included an initial attempt to identify potential uses of information and analysis from BER and related User Facilities in meeting these energy sustainability Grand Challenges,[25] At that time, it was anticipated that there would be only a very modest amount of use of information from existing Facilities in addressing the emerging challenges. Since then, two pathways have been recognized by which the impacts of the capabilities at the User Facilities could have much more significant influences

[24] Energy Frontiers Research Centers: https://science.energy.gov/bes/efrc/research/.

[25] USGCRP. 2016: https://climatemodeling.science.energy.gov/sites/default/files/Multi-Model_Framework_WorkshopReport_Dec_2016_Final_web.pdf.

in meeting the grand research challenges in energy sustainability. First, energy sustainability research relies heavily on inputs from the Earth, environmental, and biological systems researchers who in turn depend significantly on inputs from the User Facilities in their work. Second, energy systems researchers could productively incorporate information from User Facilities on advanced energy technologies directly in their modeling work, including advanced battery materials, biofuels production and conversion technologies, and projected renewable energy resource potentials.

The Grand Challenges report also included an assessment of the potential relevance of User Facility resources toward meeting two action items that were recommended in the energy sustainability chapter. Conclusions then were that there was some potential to take advantage of User Facility resources in implementing these action items.

Recommendation

5.1 Establish a strategically distributed network of research centers focusing on the Science of Energy and Environmental Resilience (SEER) that would develop and apply an array of capabilities for evaluating and projecting the dynamics of coupled human and environmental systems in support of national needs.

Science of Energy and Environmental Resilience Centers

Although there do exist a number of building blocks and nascent activities that are slowly building our understanding of energy and environmental resilience, a much more focused and ambitious effort is needed to meet this Grand Challenge. We recommend the establishment of a strategically distributed network of research centers focusing on the Science of Energy and Environmental Resilience (SEER). The SEER Centers would fill a vital national need by dramatically improving the scientific understanding of how the nation's co-evolving human and natural systems, especially those related to energy production and use, are changing across multiple spatial and temporal scales. This includes not only understanding the potential efficacy of different energy futures and their impacts on other human and environmental systems, but also how these systems influence the time-evolving feasibility and effectiveness of different energy system configurations in different regions, as well as at the national scale. Although the primary focus of the SEER Centers would be on developing a predictive understanding of the complex interactions among energy systems, other human systems, and the natural Earth system, these understandings also would contribute to several other Grand Challenges from the 2017 BERAC report.

Key science questions that the SEER Centers would address include:

- How resilient are the nation's current energy and environmental systems, individually, collectively, and in combination with other systems and sectors, and which factors or combinations of factors contribute most significantly to changes in resilience?
- How do the particular characteristics of landscape settings, including natural and built environments, interact to affect the resilience and environmental sustainability of our energy systems?
- What are the best metrics for evaluating the resilience of individual systems and collections of systems in a multisectoral context, and which data and tools need to be developed to establish and track resilience along different axes?

- Which characteristics of complex, interdependent systems contribute to instabilities, inflection points, and other nonlinear behaviors that could lead to rapid changes in resilience (either positive or negative)?
- What are the key uncertainties associated with energy and environmental resilience, and what are the best ways to understand and explore these uncertainties (e.g., through scenarios and multimodel ensemble projections)?

The multidimensional nature of energy and environmental resilience demands that the SEER Centers collectively address a broad range of research topics, many of which would build on existing research activities or programs. Like the BRCs and EFRCs, which were established "to accelerate transformative discovery, combining the talents and creativity of our national scientific workforce with a powerful new generation of tools," each SEER Center would bring together a unique combination of people and tools to focus on specific aspects of a Grand Challenge.

Some potential examples of research themes that one or more of the SEER Centers could address include:

- Energy-Water-Land Dynamics
- Human Population Dynamics and Urban System Resilience
- Coastal System Dynamics and Resilience
- Natural Resources and Material Flows
- Ecosystem Services
- Technological Innovation
- Institutions and Governance
- Complex Systems Theory and Methods

Similarly, the spatial heterogeneity of the nation's energy and environmental systems demands that the SEER Centers incorporate a distributed, federated strategy that can support context-specific analysis. This strategy is similar to the motivation for the spatially distributed BRCs, as well as ARM sites, which are strategically located to help characterize key processes and interactions that occur in specific geographic contexts. The exact regional distribution of SEER Centers is somewhat flexible and would be determined through a comprehensive selection process driven by the topical foci and science questions above, but with attention paid to regional balance. As a potential starting point, one could imagine at least one SEER Center in each of the regions defined in the National Climate Assessment (see Figure 2, below). Each Center would include a combination of partners from academia, national laboratories, industry, and the nongovernmental organization (NGO) community with expertise relevant to a Center's topical foci and regional context.

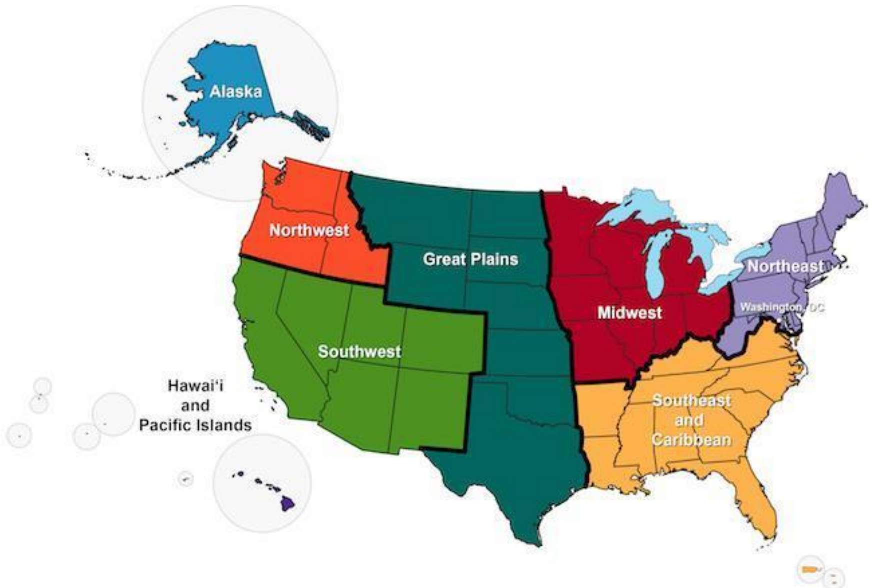

Figure 2. National Climate Assessment Regions.

Finally, each SEER Center would concentrate on the development of a unique combination of tools needed to address its topical focus areas and geographic contexts. Comprehensively evaluating the resilience of coupled energy-environment systems is a significant scientific challenge that will require the development of a wide array of advanced data, modeling, and analysis tools that can account for the extreme complexity and diversity of relevant processes and interactions. The SEER Centers would leverage DOE's world-leading expertise in developing and applying computational tools to meet these challenges, as well as capabilities funded by other sponsors that address specific topics or contexts (see examples below). The sheer diversity and complexity of systems involved will demand the development of radical new methods for bridging different systems and scales. There also will be a need for coordination and cross-fertilization of ideas across the SEER Centers, developing common metrics and ontologies, and other crosscutting activities that will help ensure the SEER program is greater than the sum of its parts.

Central User Facility

Individually and collectively, the distributed SEER Centers will develop and apply an array of capabilities for evaluating and projecting the dynamics of coupled human and environmental systems in support of national needs. However, the interconnectedness of the nation's energy and environmental systems, along with the practical aspects of coordinating and leveraging advances across the network of Centers, demands a fairly substantial degree of central coordination. In addition, there are a number of capabilities that will benefit nearly all the Centers, as well as related efforts sponsored by other DOE programs or other agencies.

Therefore, we recommend that the SEER network include a central Coordination, Integration, and Visualization (CIV) Center and User Facility that would be responsible for coordinating, synthesizing, and increasing the impact of the distributed SEER Centers. The CIV Center would serve as a hub, clearinghouse, central repository, and resource for data, analysis, and modeling capabilities and efforts at the distributed SEER Centers. These services would include, for example, data assets that benefit multiple individual Centers, efforts to develop coordinated future scenarios as well as resilience metrics and common uncertainty characterization approaches, and expertise and resources for developing repeatable and traceable model coupling approaches. The CIV Center also would be responsible for cross-fertilization of ideas, coordination of efforts to ensure diversity of approaches without duplication of effort, and capacity-building in key areas.

The CIV User Facility, which would be colocated with the CIV Center, would further advance the SEER network by providing advanced computational support and next-generation analytic visualization capability supporting both online and in-person scientific exploration by users both inside and outside the SEER Centers. Part of the mandate to the individual SEER Centers would be to develop tools for inclusion in such a facility, as part of an open-source development and application paradigm designed to dramatically increase the reach and impact of individual development efforts. CIV users initially would include mostly collaborators from other SEER Centers seeking new ideas and approaches for addressing common problems, as well as for intercomparing models. Over time, however, the SEER CIV User Facility would evolve to support scientific analysis and scientifically driven decision making by a much broader range of collaborators from academia, national laboratories, industry, and the nonprofit community.

This central "user facility" would be a place where visitors, postdocs, and students could apply to facilitate their energy and environmental resilience research, interacting with other visitors and permanent Center staff while using its database, computational, and visualization resources. Remote access and collaboration resources would also allow for virtual collaborations.

A final, important role that both the CIV Center and User Facility would play is to coordinate outreach and two-way interactions with other DOE programs, User Facilities, and activities. For example, CIV staff would maintain contact with individual DOE offices responsible for different aspects of the energy system, as well as the broader scientific community, making sure that the SEER Center analyses leverage evolving sector-specific tools and understanding (e.g., performance characteristics of batteries, life-cycle assessment of different biofuel production strategies, and urban-scale analysis tools). The CIV Center could also promote the propagation of SEER insights to inform DOE priority-setting (e.g., investments in specific research foci that lead to dramatic increases in overall system resilience across a wide range of future scenarios). This final function will be difficult but also critically important to ensuring that the SEER Centers are viewed as independent and credible sources of comprehensive analysis.

Possible Examples of SEER Centers

The following examples provide a hypothetical snapshot on what the proposed SEER Centers might focus and deliver.

Example 1: Developing a Deeper, Science-Driven Understanding of U.S. Biomass Futures

How might natural and socioeconomic resources, evolving industrial and energy systems, science and technology advances, regional markets, changing weather patterns, and land and water resources influence the evolution of biomass production systems and their innovative applications in the U.S. economy?

Motivation. A growing body of research has shown how cellulosic biomass production can contribute in several ways to making the U.S. energy system more sustainable, but most studies so far have considered the potential of biomass for individual applications. DOE's *2016 Billion-Ton Report*[26] evaluated the production potential for ethanol from biomass across the continental United States, and other studies have assessed the potential biomass production for particular purposes (e.g., electricity generation with

[26] U.S. DOE. 2016. *2016 Billion-Ton Report: Advancing Domestic Resources for a Thriving Bioeconomy, Volume 1: Economic Availability of Feedstocks.* M. H. Langholtz, B. J. Stokes, and L. M. Eaton (Leads), ORNL/TM-2016/160. Oak Ridge National Laboratory, Oak Ridge, TN. 448p. DOI: 10.2172/1271651 (http://energy.gov/eere/bioenergy/2016-billion-ton-report).

carbon capture and storage[27]). These use(s) of biomass production are likely to vary spatially across the country, depending on the kinds and amounts of biomass that could be produced and the constraints on storage, processing, and transport, as well as other factors. Decision makers should also consider the broader carbon cycle implications of a massive switch to biomass production, including not only displacement of fossil energy sources, but also fundamental changes in carbon storage and fluxes across large areas that could become significant to atmospheric CO_2 levels.

Approach. Integrated and coordinated campaigns entailing experiments, data collection, and modeling are needed to develop spatially explicit, comparative scenarios of the potential for biomass production systems to contribute to environmental sustainability, broadly defined, in diverse regions of the country. At least three focal study areas could be established in regions with potential for high biomass production (including dedicated energy crops as well as agricultural residues) but variable potential for transport and processing of biomass or bioproducts. The four existing BRCs could contribute data and scientists for these campaigns.

The following are specific examples of how biomass could contribute to our national energy portfolio and/or mitigation of impacts of fossil energy sources.

- Biomass production can offset some of our dependence on petroleum, as exemplified by the U.S. bioethanol industry.
- Ethanol production based on cellulosic biomass grown on marginal lands can reduce the need to produce corn grain ethanol, thereby reducing use of cropland for fuel instead of food.
- Biomass production can produce feedstocks to substitute for petrochemicals currently used to produce specialty biofuels and bioproducts (e.g., isobutanol).
- Biomass can be co-fired with coal in electricity generation (see below: Electricity-Biomass-Agriculture Interactions).
- Biomass production can be used in bioenergy with carbon capture and storage (BECCS) technology, which is one of the most promising approaches to achieve negative emissions, thus helping to mitigate undesirable changes in atmospheric composition. BECCS is most feasible where lands with potential for high biomass production overlap with suitable geological features for injection and storage of CO_2, or pipeline transport capabilities to carry CO_2 to such features.[28]

Outcomes. The campaigns would produce comparative analyses, including technoeconomic and life-cycle analyses, at nested spatial scales from counties to regions. Results would improve our understanding of (1) how human activities interact with, and increasingly perturb, the carbon cycle at regional to global scales and (2) how such perturbations could be better managed in the future. This research program would facilitate the technical and economic comparisons of potential use of biomass with energy production from fossil fuels, as well as with other renewable energy sources such as wind and solar. They also would provide a basis for comparison of the potential environmental benefits and harms of the alternatives. Results are likely to differ by region. Gaps in knowledge requiring further research would be revealed.

---

[27] Baik, E., et al. 2018. "Geospatial Analysis of Near-Term Potential for Carbon-Negative Bioenergy in the United States." In *Proceedings of the National Academy of Sciences USA* **115**(13), 3290–95. DOI:10.1073/pnas.1720338115.
[28] Baik, E., et al. 2018. "Geospatial Analysis of Near-Term Potential for Carbon-Negative Bioenergy in the United States." In *Proceedings of the National Academy of Sciences USA* **115**(13), 3290–95. DOI:10.1073/pnas.1720338115.

Example 2: Resiliency in Energy and Environmental Systems from Local to Regional Scales: The Northeast Region

Motivation. The Northeast is a compelling geographic context to advance scientific understanding of the complex interactions, interdependencies, and co-evolutionary dynamics of energy and environmental systems at local to regional scales under multiple stressors. Making this region a viable context are its extensive natural gas resources and pipeline network, an electric transmission system with significant congestion, an electric generator fleet heavily dependent on once-through cooling, and conflicting multisectoral demands on the water and land resources. This region is home to many urban centers confronting multiple infrastructure stressors that are both systemic (e.g., aging infrastructure) and acute (e.g., heat waves and storm-induced flooding). Individually, and especially in combination, these urban and regional contexts provide a rich testbed for studying multiscale energy-water-land interactions, including potential tipping points and cascading failures under complex stress conditions.

Approach. The approach would entail a data collection initiative working closely with utilities and local or regional agencies. Also necessary would be data reconciliation and translational tools to both aggregate and disaggregate existing data to match the necessary resolution for modeling. Modeling capabilities and tools would need to span multiple scales (i.e., regional and local), both demands and supplies as well as markets and trade, and institutional barriers. For example, the necessary electricity models would have to span generation, transmission, and distributions on the supply side, and electricity demands by industry, buildings, and transportation end-users on the demand side, as well as how both supplies and demands including markets and prices may fluctuate under different human or environmental influences. Similarly, water demand and supply models would be essential to model the evolution over time of water demands and supplies (i.e., runoff, groundwater recharge, snowpack, and reservoir storage) and how those changes affect the evolution of the electricity fleet, bioenergy and cropland expansion. At the urban scale, these modeling needs also would entail modeling the interactions among the water distribution system, power distribution system, natural gas network, transportation network, and the transitional tools necessary to facilitate the exchange of information among these models at the appropriate temporal and spatial scales.

Outcomes. The testbed can establish an array of different data, modeling, and analytical tools to understand the risks and resilience of individual and connected systems as well as the consequences of strategies to enhance resilience at both local and regional scales. Impactful and relevant simulation and analysis require access to high-quality data at the appropriate spatial and temporal scales to calibrate and validate modeling and analytical tools. Establishing such data products and identifying the most appropriate types and aggregation levels will inform data-collection initiatives. Moreover, establishing open-source modeling capabilities and tools that adequately span the sectors, processes, and temporal and spatial scales that are needed to address energy-water-land interactions can transition to becoming community-wide modeling capabilities, which can be applied in other geographic contexts. The testbed would bridge an important gap in understanding how energy-water-land systems evolve and interact across multiple scales and in advancing the data and tools necessary to co-manage these systems and ensure a sustainable future energy system. Many of the lessons learned and tools developed to study the Northeast testbed would be applicable to other geographic contexts and shared with other Centers through close coordination with the CIV Center.

Example 3: Electricity-Biomass-Agriculture-Water Interactions

Motivation. The co-firing of existing coal plants with biomass has been considered a strategy to extend the economic life of these plants. Increasing biomass co-firing will require an expansion of the supply of bioenergy feedstocks beyond the current use of forest residues to other sources of supply such as crop residues.

Key research questions include: If biomass co-firing for coal units were adopted at a larger scale, what are the impacts on the coupled energy-food-water system? For example, where would the bioenergy feedstocks come from and how much would be demanded by these coal units? What are the impacts on agriculture and water for irrigation? What are the impacts on nitrate leaching? What are the impacts on air quality?

Approach. Capturing these electricity-biomass-agriculture-water interactions requires the coupling of sector-specific models and data to enable physical and economic feedbacks. This coupled system would include a power system model for the power grid region that links unit commitment models for individual power plants together into an integrated market for electricity, augmented to allow the co-firing of coal with biomass at a prespecified level.

Power plant–specific, biomass supply curves would be incorporated into this coupled system, based on the spatial distribution and density of feedstocks, including forest residues and corn stover, enabling the endogenous determination of biomass supply costs. Where power plants draw on the same biomass supply shed, the competition among power plants and with other prospective biomass markets will need to be represented. This would capture the fact that one plant's decision to co-fire biomass will depend on the other plant's co-firing decisions, as well as competition in the electricity market within the region.

The coupled system would include a gridded model of crop production. Because the bulk of U.S. corn production falls within the Midwest region, this model would need to endogenously determine the spatial and economic response of corn production to co-firing decisions. This gridded crop model would be linked to the power system model through the premia paid for corn stover—the dominant biomass feedstock in this region. Corn producers falling within the biomass supply shed for a co-firing power plant respond to these by-product payments by increasing corn crop area and intensifying production. This, in turn, has consequences for local water quality, since nitrate leaching is the main source of water quality degradation in the region.

To bring in potential weather impacts on this energy-land-water system, models would need to incorporate estimated grid cell–specific yield response functions for rainfed and irrigated corn and soy production within the region. These Earth system impact estimates could be improved by including historical soil moisture estimates generated by a water balance hydrology model in place of precipitation, which is but one input to the critical soil moisture index.

Using global gridded climate model outputs, estimates of crop yield level, variability and sequencing under current and future climatology could be generated, allowing for an assessment of impacts to irrigation under future changes in the Earth system and raising questions of future irrigation demand in the region and hence groundwater sustainability. These questions necessitate the addition of a water balance model to this coupled system.

Outcomes. This type of fine-scale integrated analysis is important for a number of reasons:

- Absent fine-scale analysis, these impacts would not be evident: Spatial competition for biomass, induced changes in land use and intensification, and induced increases in nitrate leaching and impacts on water quantity and quality.
- This framework offers a means of identifying and quantifying tradeoffs among energy, land, food, and water objectives.

Conclusions

Collectively, the SEER Centers and CIV would dramatically enhance DOE's ability to meet its mission to ensure America's security and prosperity by addressing its energy and environmental challenges through transformative science and technology solutions. Individually and collectively, they would offer the following opportunities:

- Provide an integrated, overarching view of the connections between the energy system and other key human-Earth systems.
- Provide experience-based information regarding desirable levels of spatial and temporal aggregation for data and analyses of integrated Earth-water-land systems.
- Develop and apply tools to assess the potential value of specific energy technologies (e.g., life-cycle assessment [LCA] and global change assessment model [GCAM] scenarios).
- Enhance capabilities for assessing the resilience of current and future energy systems to future human-Earth system changes.
- Improve projections of net biogeochemical fluxes.
- Characterize the most promising paths for energy sustainability; multifacility management of the Network of Energy Sustainability Testbeds (NEST) and User Facilities can inform investments and activities at the other User Facilities.
- Lay the foundation for a future User Facility (hub) that provides broader community access to these foundational capabilities.

Challenges include:

- Relationship to current activities in Multisector Dynamics (e.g., Integrating Human and Earth System Dynamics [iHESD]; Program on Coupled Human and Earth Systems [PCHES]; Integrated Multisector, Multiscale Modeling [IM3]; and others that include some dimensions of energy sustainability); mechanisms are needed for cross-fertilization, as well as periodic idea sharing.
- Data for energy systems, including access and harmonization.
- Handling proprietary data, including rules and procedures for access.
- Leveraging the Earth System Grid Federation (ESGF), Program for Climate Model Diagnosis and Intercomparison (PCMDI), and Global Trade Analysis Project (GTAP).
- Accounting for complex human system dynamics, including legal constraints, institutional paradigms, agent-based modeling.

Chapter 6. Computation and Data Analysis

The 2017 BER Grand Challenge Report[29] identified data analytics and computing as one of the major focus areas for BER over the next 20 years, a crosscutting theme that also arises in different forms in the other topic areas. The articulated vision for data analytics and computing is very broad:

> *Develop the approaches and computational capabilities to collect, store, and analyze large-scale data across temporal and spatial scales.*

CHARGE 1 RESPONSE

Alignment of User Facilities to Current BER Computation and Data Analysis Needs

DOE's User Facilities have the opportunity to coordinate efforts and provide the research community with an interconnected infrastructure for simulation as well as archiving, managing, analyzing, and visualizing experimental, observational, and model data, and metadata. Such an infrastructure will support the integration and management of models, experiments, and data across a hierarchy of scales and complexity to accelerate the pace of scientific discovery and predictive understanding of the Earth system.

Some of the large BER projects, such as the Earth Systems Grid (ESG) and DOE's Systems Biology Knowledgebase (KBase) project, manage their own hardware for data storage and computing hardware, but the most significant volume of both comes from BER and ASCR facilities. The following BER facilities have integrated computing and data capabilities:

- ARM currently has two compute clusters for modeling, analytics and learning. With 1,150 users, ARM spends roughly 20% of its operating budget on data product development and data management, which include extensive metadata tracking, making searching and selecting specific datasets easier to download or analyze.
- JGI has over 7 petabytes (PB) of storage, divided into multiple file systems optimized for different usage models and has also developed data management software (JAMO) and a community metatdata service (IMG-GOLD), as well as topic-specific databases (e.g., IMG). JGI runs its own large-memory servers and has also partnered with NERSC to acquire one rack of its Cori supercomputer, with a total of 6K Haswell cores. This rack provides a guaranteed allocation of JGI time and a separate queue to access the system, but it leverages some of the economies of scale from running the larger system. JGI leverages ASCR facilities for long-term preservation of data in their tape archives.
- EMSL has a single large-cluster 1,440-node Intel Linux cluster (Cascade), which has two Xeon Phi co-processors attached to each 16-core Xeon CPU. EMSL also runs its own data storage archive (currently ~2.5 PB of disk space and 14.5 PB of tape archiving). It provides a community data repository (MyEMSL) for sharing data and visual analytics tools to interact with and explore data.

[29] BERAC. 2017. *Grand Challenges for Biological and Environmental Research: Progress and Future Vision; A Report from the Biological and Environmental Research Advisory Committee*, DOE/SC-0190, BERAC Subcommittee on Grand Research Challenges for Biological and Environmental Research (science.energy.gov/~/media/ber/berac/pdf/Reports/ BERAC-2017-Grand-Challenges-Report.pdf).

ASCR facilities also provide computing and storage services to BER researchers:

- NERSC currently operates two supercomputers, which will deliver over 9 billion hours of computing time in 2018 to over 7,000 users in the Office of Science. One of those systems, Cori, is a Cray XC40 with two different kinds of nodes, 2,388 Intel Xeon "Haswell" nodes with 32 cores each and 9,688 Intel Xeon Phi “Knight’s Landing” nodes with 68 cores each. Cori also has a 28-PB disk storage system as well as 1.8 PB of Solid State Disk (SSD) that operates at 1.7 terabytes (TB) per second for data-intensive application. NERSC has a second system, Edison, which will be replaced by one that is more powerful than Cori by 2020. NERSC has a tape archive with over 120 PB of data, which grows by 170% per year and includes over two decades of scientific data. The archive also has multiple disk-based file systems that are optimized for different usage needs; there are quotas on disk usage, but projects with large demands for online storage can purchase additional space at NERSC. NERSC also operates a variety of databases and “science gateways” to serve some of the largest NERSC datasets to the broader science community. NERSC allocations are directly managed by DOE programs, so BER can manage the pool of time based on program research priorities.
- Leadership Computing Facilities (LCFs) at Argonne National Laboratory and ORNL (ALCF and OLCF, respectively) provide significant computing to BER. OLCF currently houses Summit—the fastest open science computer in the world at over 187 peak PETAFLOPS. Summit is an IBM system with Power9 CPUs, NVIDIA GPUs, and an Infiniband network. It is currently in preproduction, so formal allocations have not begun, but OLCF also has a production system, Titan, at over 27 peak PETAFLOPS. Titan is a Cray XK7 system with advanced micro device (AMD) CPUs and NVIDIA Kepler GPUs. ALCF now operates two main supercomputers—Mira (an IBM Blue Gene/Q system) and Theta (a Cray XC40 system with Xeon Phi nodes). Both systems are using lightweight processing cores for energy-efficient high performance and for custom high-performance computing (HPC) networks from IBM and Cray, respectively, for scalable application performance. At around 10 PETAFLOPS each, ALCF allocates more than 5 billion computing hours each year to over 1,000 users. Both LCFs plan to install exascale computers in the next few years. ALCF has over 40 PB of storage associated with its supercomputers and 65 PB of archival tape storage. OLCF will soon have 250 PB of storage associated with the Summit system, as well as archival storage. The data-retention policies prescribe keeping data for 90 days after the end of a project. The majority of allocations at both facilities are managed through the INCITE process, an annual peer-reviewed process that is open to the broad national and international scientific communities in academia, industry, and the national laboratories; it is designed to meet the needs of the most computationally demanding scientific challenges, whether they involve simulation or data analysis. In 2018, for example, BER-related activities received close to 1 billion computational hours in allocation, with the bulk in Earth systems and geology. In addition, the ASCR Leadership Computing Challenge is a separate annual peer-review process that allocates resources at ALCF, OLCF, and NERSC to projects with an emphasis on high-risk, high-payoff simulations in areas directly related to the DOE mission and for broadening the community of researchers capable of using leadership computing resources.
- ESnet is the wide area network connecting all the national laboratories to each other and to the rest of scientific community via the internet. Unlike commercial providers, ESnet is engineered to support high bandwidth for enormous data transfers, which is necessary to move scientific data between sites. This network includes support for bandwidth reservations, continual monitoring, and upgrades of links that are close to saturation, as well as monitoring tools to identify problems. ESnet is planning a major upgrade to increase the bandwidth to 1 terabit per second in the next few years.

The existing ASCR facilities provide a substantial resource for BER's computing and data storage needs, with some support for software and training. BER facilities as well as some of the large BER projects provide additional support for software and data services specialized to BER needs. Taken as a whole, they provide a range of services and infrastructure, but there are still gaps relative to the needs of the research program.

Alignment of Existing Facilities to BER Computing Needs

The large ASCR computing facilities have many aspects that are well aligned with the requirements identified in the 2017 Grand Challenges report.[30] They provide (1) compute cycles on next-generation (exascale and pre-exascale) architectures; (2) training and support for adapting codes and algorithms for these new machines; and (3) installation, coordination, and user support for some large community codes. With respect to computing capability, the LCF facilities are deploying pre-exascale systems and are on track to deliver at least one exascale system as early as 2021. The LCF systems are designed for capability simulations—simulations that require a large portion of the machine and are difficult, if not impossible, to perform elsewhere. This requirement means these systems support a small number of users, and thus they can support only a small number of BER science applications. NERSC is also deploying pre-exascale systems, but it serves a much larger user community with over 7,000 users and roughly 700 different application codes that run at a wide range of scales. The LCFs and NERSC are also well aligned with respect to education and training and support of community codes. These computing centers run frequent training workshops and hackathons. They install some of the most popular community codes in a way that is tuned to user platforms and can assist users in some configuration and use issues. They also have early access programs for planned hardware upgrades: NERSC's NESAP, OLCF's CAAR, and ALCF's ESP. These programs provide access to prototype hardware, workshops with vendor experts on optimization, and support for postdocs who work directly with application scientists.

The existing computing facilities are not well aligned with Grand Challenge needs in several ways. They include the desire for faster turnaround, especially for midrange computing, and more access to traditional general-purpose processors.

Job Turnaround

Many of BER's high-end computing users are served by NERSC, with over 700 total projects, each serving the science needs of the principal investigator(s) and their team. NERSC is heavily allocated and heavily used; in fact, a utilization of over 90% was reported during our workshop breakout session. However, such high utilization and the wide range of job sizes sometimes translate into long job turnaround times, resulting in a very efficient resource usage but often a wait of several days in the queue before obtaining 1 day of execution time. This long wait time causes a large reduction in the effective throughput rate from the user perspective. We compared this process to cloud computing, which reportedly often runs at well under 50% utilization and main support jobs with small node counts to have a high probability of on-demand access. We also noted that many applications have purchased additional resources to obtain near on-demand access, with JGI purchasing hardware at NERSC, which they run at 80% utilization. EMSL noted that its facility runs at 85% utilization with a more uniform jobs mix, resulting in a dramatic reduction in turnaround times.

---

[30] BERAC. 2017. *Grand Challenges for Biological and Environmental Research: Progress and Future Vision; A Report from the Biological and Environmental Research Advisory Committee*, DOE/SC-0190, BERAC Subcommittee on Grand Research Challenges for Biological and Environmental Research (science.energy.gov/~/media/ber/berac/pdf/Reports/ BERAC-2017-Grand-Challenges-Report.pdf).

Real-Time Computing

The high utilization makes difficult the fast-turnaround, "human-in-the-loop" type computing, also prohibiting real-time processing. Responding to this demand, NERSC runs an interactive job queue for jobs with modest node counts (64 or fewer) in under 4 hours. It also has created a pilot program for real-time jobs, with Cryo-EM analysis as one of the initial pilots. But the current allocation mechanisms and policies do not guarantee real-time access for midrange or large jobs, so application communities may need to purchase their own resources, either within one of the ASCR systems or as a stand-alone facility.

General-Purpose Processors. We also identified a need for more access to general-purpose processors. NERSC and the LCFs are well aligned with providing next-generation processors and other cutting-edge technologies to continue growing computing capability to match the needs of users. These new processors promise better performance while using less power, but they may require large investments in software development. Some codes can adapt quickly to these architectures. Larger codes may require several years of development before they can be run effectively. To aid in this transition, ASCR's funding of the Exascale Computing Project (ECP) includes over $2.5 million per year for moving codes like E3SM to the future exascale platforms. Some codes, especially those used in a lot of midrange computing, still run better on conventional general-purpose CPUs. This may be both an inherent mismatch between the kinds of fine-grained parallelism required in the advanced architectures or insufficient motivation and funding to rewrite the codes. The demand for general-purpose CPUs is evident when comparing the job backlog of the Cori Haswell system (general-purpose CPUs) and the Cori Xeon Phi system (prototype exascale architecture). To obtain additional midrange computing resources, many applications purchase their own cycles or hardware, and these machines are almost always general-purpose CPUs: EMSL is running a 1,440-node Intel Linux cluster; although much of the theoretical peak performance comes from two Xeon Phi co-processors on each node, the large majority of EMSL user jobs do not use the coprocessors but rely heavily on the 16 conventional Xeon cores available on each node. JGI purchased its own Xeon Haswell nodes for installation at NERSC. The E3SM project purchases 240 nodes (Xeon Broadwell), and the ARM project runs two clusters (30 and 112 nodes) of conventional general-purpose processors. While this preference for general-purpose high-end processors is clear, there have been exceptions. JGI has used FPGAs for some of its genomics processing, because these codes tend to operate on 2-bit or 4-bit words and do not use floating points. For many image-analysis tasks, GPUs have proven very effective and are the preferred architecture for the BES-ASCR CAMERA project, for example, which develops algorithms and implementation for the light sources and other facilities. Notably, one of the fastest computations in the world, running on the Summit computer at ORNL, is a biology application at over 1.8 Exaops using 16-bit arithmetic, and a second deep-learning example on Summit, also using 16-bit arithmetic, finds extreme weather events in climate simulation data.

Support for Complex Workflows

Many BER workflows are becoming quite complex, requiring many different applications, compute and memory resources that vary across the workflow, and control of the software environment (e.g., older versions of compilers). In addition, the computing challenges are not entirely separable from the data challenges described below, as complex simulation workflows may require multiple datasets from remote sites and thus the ability to reserve network bandwidth as well as computing resources simultaneously. Support for these workflows may be easier for a facility with a singular mission of supporting a few midrange production applications, although at a higher cost than facilities with a larger user base and computation system, over which support costs can be amortized. Containerized software helps with some of the software issues, including NERSC's Shifter software for containers on their HPC systems, as well as support for interactive jobs (e.g., launched from Jupyter notebooks).

Other Issues

BER already is investing in dedicated computing at the midrange but on a project-by-project basis. ARM, EMSL, JGI, E3SM, KBase, and probably several smaller projects run their own dedicated resources. Coordinating these efforts within BER to create a single consolidated facility would result in increased efficiency and the ability to share common resources and jointly address common challenges such as long-term software curation and workflow tools for data-model integration. These BER computing facilities would also be a natural place to build large-domain expertise for user support for BER critical applications, more than can be provided at the ASCR facilities that serve a broad science community.

In summary, while BER computing requirements will mostly be met at existing BER or ASCR facilities, some changes may be required in allocations, scheduling policies, and additional funding to ensure the desired access:

- Fast turnaround.
- Better real-time support and interactive computing on midrange jobs.
- General-purpose processors for codes that have not adapted to advanced architectures.
- Complex workflows (software interfaces) for data-model integration.
- Coordination across facilities to ensure uninterrupted access for BER projects.
- Long-term software maintenance.
- Additional domain expertise for user support.

Alignment of Existing Facilities to BER Data Needs

Data challenges in BER research programs have increased by orders of magnitude over the past few years. Common approaches can be employed among BER programs for archiving, accessing, processing, and generating enhanced data products. Given that data volumes already exceed exabyte scales and that some of the community datasets are produced and supported by multiple governmental agencies, a federated data system approach is needed for future expansion. This expansion will require that data providers maintain a set of geographically dispersed sites accessible by the scientific community and that, through linkages, nodes increasingly will be required to serve as archived data repositories, thus allowing scientists to access all data as if they were on their own system.

The BER and ASCR facilities such as EMSL, NERSC, OLCF, and ALCF have tape archives that provide persistent storage of data and low overall cost to DOE, enabled by economies of scale across programs. Science gateways and other means of supporting domain-specific portals have been used effectively at NERSC, JGI, and other facilities to support customized access to data and computation for large communities of users. These portals reduce the barrier to entry for new users and enable “curated” views of available resources and ready access to appropriate tools. Linkages among federated system members are provided by ESnet with data-driven science features that provide expedited data transfer capabilities. Cross-network coordinated bandwidth reservations enable high-speed data transfer among institutions involved in BER science. Data Transfer Nodes and Science DMZs provide a local end-point at each institution for transferring data (adopted standard by DOE, NSF, and other agencies). The planned upgrade of ESnet to terabit by 2022 will continue to expand this capability, and plans for advanced data services will enable easier, automatic management of the network resources required to move large scientific datasets across sites.

Alignment of Existing Facilities to BER Software Needs

The development and maintenance of software continue to pose significant challenges to the scientific enterprise. The BER User Facilities all have significant computational elements that require the use of community software as well as development of custom software. BER programs and User Facilities have made significant investments in large community modeling frameworks such as NWChem (computational chemistry, EMSL), E3SM (Earth systems modeling, CESD), and PyART (weather radar models, ARM). The installation and execution of codes needed by individual users at NERSC and other computational facilities have been facilitated by the development of container technologies. However, several challenges to effective development, maintenance, and use of software at BER User Facilities were identified as follows:

1. There is little coordination of software practices across the User Facilities. Where there exist complementary research activities (e.g., omics studies at both JGI and EMSL requiring bioinformatics software), each institution typically develops and uses its own software and data analysis workflows. Some coordination may occur at the project level, such as through FICUS projects, but, in general, there may be opportunities for improved coordination or collaboration in community software development that could benefit multiple facility user communities. Code sharing and community development could be enhanced through the use of emerging community resources such as the "DOE Code" GitHUB repository managed by the Office of Scientific and Technical Information (OSTI; .www.osti.gov/doecode/).
2. Although some software development is directly funded by the research program and some level of support (e.g., installation, tuning, and reporting user issues) by User Facility budgets, in neither case is there a commitment of long-term funds for software maintenance, documentation, training, and upgrades.
3. Because of limited budgets for development, the perceived urgency of software needs, and coding by domain scientists, in many cases rigorous software engineering practices are not followed. This, combined with the above challenge 2, often leads to the production of software that is not well documented, not reusable for related needs, lacks interoperability with other user software, and is difficult to maintain or upgrade.
4. BER user science is commonly interdisciplinary in nature and draws on multiple modes of instrumentation, combinations of several different analysis and modeling techniques, and integration of multiple data streams. Other than a few regularly repeated standard analysis workflows, there is little support for the complex modeling and data analysis workflows that are needed to enable this type of science. An opportunity exists to address this challenge through development of a flexible network of well-designed analysis and simulation modules, linked through standardized interfaces and supported by resident domain experts that could be efficiently configured to meet the needs of individual users. These linkages might be multiphysics (e.g., linking biological, hydrological, and geochemical models), multiscale (e.g., linking models with different levels of spatial, temporal, or process fidelity), and/or multimodal (e.g., linking data from different types of instruments).

Alignment of Existing Facilities to BER Needs in Training and Support

New techniques and services are required to leverage the wealth of research results and transform them into world-leading scientific discoveries. Workforce development will enable BER to take full advantage of these advancements.

The BER, BES, and HPC User Facilities all provide user training in the form of tutorials and user meetings. This training provides the vital knowledge users need to access and make use of the facilities. The facilities also have groups dedicated to helping the users with special needs (e.g., porting of codes to computational machines, special instrumentation, and additional measurements). This training and assistance represent an important component of the success of these User Facilities.

The benefits of the new emerging data and computation capabilities envisioned in this report require a workforce that is ready to take advantage of opportunities. Following is a list of some examples of required skills:

- Create data that can be integrated with other data available
- Utilize available remote computational resources
- Generate data for re-use by others unfamiliar with the data generation process
- Utilize advanced modeling, analysis, and visualization tools
- Leverage emerging data science tools

Focused workforce development programs, including multiday-per-week concentrated courses and ongoing partnerships with emerging data science programs at universities, would help ensure that DOE data and computational resources continue to yield benefit beyond their initial collection. In addition, the creation of a program of "*campus champions"* (a term borrowed from NSF's XSEDE program) could help create new expertise in computing and data technologies and resources available within DOE facilities and the broader scientific community.

CHARGE 2 RESPONSE

Alignment of User Facilities to Address Future Needs and Grand Challenges in Computation and Data Analysis

The Grand Challenges report lists five challenges in the area of computation and data analysis. Several other Grand Challenges require significant support of computation and data facilities.

Grand Challenge 6.1: Develop robust approaches for large-scale data collection, curation, annotation, and maintenance.

Many of the data challenges are cross-facility issues where data are collected at one facility or an observational site, transferred using ESnet or other means, stored at another facility or lab, and possibly served by yet another facility. The responsibility for curation, annotation, and maintenance often falls to individual investigators and are *ad hoc*, if they are done at all. In addition, many of the datasets used in BER science challenges will use derived data products, such as reconstructed Earth systems data, postprocessed simulations, assembled genomes, or extracted genes. Scientists must be able to trust these derived products and understand the limits of confidence, error bars, and other quantitative metrics. Thus, along with data collection, data need to be labeled and metadata analyzed to allow data of particular interest to be located, but also reprocessed as improved methods are discovered. A priority in this area is to make the data reusable and scientific experiments repeatable through inclusion of sufficient metadata and the preservation of raw data and tools to reproduce particular versions of datasets.

Science Focus Areas (SFAs), such as the Watershed Function SFA, require the combined analysis across datasets, and thus bringing diverse data from a variety of sources and types into an infrastructure with common data formats and vocabulary. Data labeling is both a social and computational challenge, requiring training of students and postdoctoral scientists, who often are directly involved in data collection, to provide consistent and meaningful labels on data. In addition, automatic metadata labeling based on the location, source, collection time, similarity to other datasets, or other features should be explored to provide consistent data labeling.

Recommendations

6.1 Provide tools at facilities for labeling, metadata management, and data discovery both within one facility and across DOE and non-DOE facilities.

6.2 Provide tools at facilities to manage derived data products, as well as long-term storage of raw data whenever possible.

Grand Challenge 6.2: Develop computing and software infrastructure to enable large-scale data (i.e., Big Data) storage and analysis.

The high-end computational facilities in DOE have traditionally focused on the modeling and simulation workload, including BER's activities in climate modeling, subsurface, environment, and molecular dynamics. The growing datasets from sequencers, electron microscopes, satellite imaging, light sources, and other instruments, in addition to massive simulation output, have created an enormous volume of data for BER researchers. Furthermore, these datasets are full of errors, have a low signal-to-noise ratio, and may involve multiple modalities, making the analysis problems very complex with analysis algorithms still under development. Analysis of these datasets is essential to many of the BERAC Grand Challenges, whether linking genotype and phenotype data in systems biology, using environmental measurements to validate and improve simulations in Earth and environmental science, or integrating molecular and process data to improve understanding of microbiome dynamics.

The computational facilities in ASCR have historically focused on modeling and simulation workloads, although next-generation systems, including exascale systems, are also including machine learning and data analytics applications in their procurement benchmarks and early science applications. Large-scale analysis problems will likely be met by these facilities, but daily production workloads from observational data, experiments, and simulations may not be. Although there is a deep bench of expertise across the DOE complex in modeling and simulation and understanding how the algorithms and application needs drive computing facility requirements, the expertise in data analysis and thus the understanding of how to exploit various computational platforms are still emerging. The processor architectures, storage systems, and networks need to be designed to serve these workloads. Even within BER, the analysis of genomic data, images, and climate simulations may require very different approaches or architectures. As described above, the data pipelines may touch multiple facilities, so analysis, data cleaning, compression, and annotation may happen on site during an experiment or at a centralized data or computing facility, or both. BER needs access to and policies for long-term data storage rather than annual allocations, as well as support for searching large distributed datasets and sharing them with the community.

Recommendation

6.3 Develop an infrastructure strategy that addresses data analysis and storage needs.

Such a strategy may combine (1) BER-managed allocations of time at NERSC, (2) LCF resources for some of the largest analysis problems, (3) commercial cloud platforms, (4) augmentation of existing BER projects or ASCR facilities with computing and storage dedicated to BER projects, and (5) deployment of *in situ* computation at major experimental sites. This approach should take into consideration cost, efficiency, and usability issues and develop approaches that work across science areas, tailoring when necessary. The strategy may be divided roughly by the scale of the data analysis challenge, such as (1) major instruments should be accompanied by their own data and computing plan, as is done with physics experiments and the climate simulation data in ESG, and (2) community repositories are needed, along with tools and standards for data that may be modest in size individually, but large scale when taken in aggregate, as is the case for genomic data. Having individuals manage these data is neither cost-effective nor in the best interest of reproducible science, and BER needs to develop a strategy that is aligned with the rest of the Office of Science, setting researcher expectations and funding priorities to address these infrastructure needs.

Grand Challenge 6.3: Conduct research to develop suitable algorithms and programming models that can harness current and future computer architectures to effectively model complex coupled systems and analyze extreme-scale data.

BER scientists will need more computing power for the complex simulations and analytics problems as part of the Grand Challenge science problems. However, with traditional technology growth stalling, any increases will come from additional parallelism and various types of specialized or “manycore” architectures. This constraint creates a natural tension between access to (1) traditional processor architectures that are familiar to programmers and tend to run software developed over decades of work and (2) architectures that can provide more computing capability and better energy and cost performance but require substantial software development. Funding should be targeted to advanced architectures, and BER-managed computing resources should be scoped to meet the need for access to traditional processor architectures. This is already happening for some of the major codes such as E3SM, and BER’s facilities provide model levels of computing for the latter. NERSC is likely to continue providing some access to traditional architectures over the next 5 to 10 years but without substantial growth, because the cost performance of such systems is not improving significantly.

Architectures that use narrower data types or have simple forms of parallelism may provide opportunities for some BER applications in genomics or analytics using deep-learning algorithms, but they have a smaller community of interest. The Exascale Computing Project (ECP) is targeting architectures that will serve a broad simulation and analytics workload and comprise application projects in climate modeling (E3SM), chemistry simulation (NWChem), subsurface modeling, and metagenome analysis codes to develop new applications or versions of existing applications that will run well on future systems. But this is far from a complete set of software needed for tackling BER Grand Challenges.

Current ASCR User Facilities have programs to help users adapt to future architectures, as well as a broader set of training programs for users. In addition, while the LCFs are focused primarily on exascale and the architectural changes necessary to meet the requirements of the largest application, NERSC has provided multiple architectures including the Haswell partition in the current Cori supercomputer. All three ASCR facilities have also provided access to testbeds, and through the SciDAC partnership program, as well as ECP, ASCR has provided partial support for transitioning codes.

A secondary concern regarding the computing facilities is the long time required for job turnaround, which can be several hours or even days for larger tasks. This is affected by prioritization of larger jobs and how heavily utilized the systems are. In addition, the need for complex data analysis workflows with long running times and modest nodes are often poorly suited to the scheduling requirements of the ASCR facilities. Several BER projects, including ARM, E3SM, KBase, and EMSL, have purchased their own computing systems to address these issues. JGI also purchased racks of the Cori Haswell system at NERSC and run that system at under 80% utilization relative to the NERSC normal of over 90% utilization.

Recommendations

6.4 Analyze the most important applications and determine which ones will need significant performance increases to meet scientific demands and which can continue with only modest increases.

6.5 Work with ASCR to ensure continued access to testbeds with emerging architectures, as well as to training and user programs to help with the evaluation and code transitioning of critical BER applications, among others.

Grand Challenge 6.4: Engineer advanced computational modeling combined with data integration across temporal and spatial scales.

New science questions will continue to drive the need for additional computational capabilities, including advances in mathematical modeling and algorithms, as well as advanced computational platforms. More sophisticated modeling techniques tend to act as more complex computational workloads, exploiting sparsity, hierarchy, and adaptivity, all of which lead to less regular and less uniform computational patterns. The result is a requirement for computational systems that can effectively handle irregular memory access patterns and interprocessor communication. Moreover, scientific simulations are no longer stand-alone batch-processed computations that run for hours based on a single set of inputs; however, they may integrate data for component models or adapt the direction of the computations. Rather than separating data and computing facilities, the two need to be supported in integrated facilities that support complex workflows that involve changing computational scales, have the ability to incorporate data from external databases, and can perform *in situ* analytics to steer computations midstream.

Recommendation

6.6 Work with the research community and computational facilities to determine the hardware, software, and usage policies needed to support researchers' complex workflows.

Grand Challenge 6.5: Conduct research and develop activities that support human understanding of large-scale, multimodal data streams, including the ability to steer experiments in real time.

Fast turnaround and predictable running times are a priority for all scientists, but experiments that use computational feedback require hard real-time constraints on the hardware and software for data transfers, analytics, and control. For this reason, some science scenarios may require substantial computation and (at least temporary) storage on site, place enormous demands on the network to transfer data, and on-demand scheduling of remote real-time computation. Further complications arise from multimodal data streaming from a variety of measurement devices and sites. In addition, both automatic steering and human-in-the-loop experiments will require new interactive simulation, analysis, and control, again breaking traditional models of batch-scheduled HPC systems.

Recommendation

6.7 Address the needs of real-time streaming data and interactive computing as part of the recommended infrastructure strategy.

Challenges for User Facilities

There are two overarching problems for the computing and data facilities, the first being the end of traditional performance scaling from computer hardware, and the second related to enabling effective scientific discovery in an era of extreme-scale data. The data challenges lie in the four Vs of big data: volume, velocity, variety, and veracity. Big data volume refers to the scale of data; velocity connotes the analysis of streaming data from experiments or simulations; variety denotes the different forms of data sometimes combined in a single analysis; and veracity refers to the noise, uncertainty, and quality issues of data. The BER research programs deal with these issues in data collection, curation, sharing, annotation, and maintenance. We identify the following eight challenges: five specific to data, one to computing, and two more that span data and compute needs.

1. **Coherent metadata is required to enable data discoverability and availability.** BER research programs generate, collect, and curate an extremely massive, dynamic landscape of information and data, and navigating through it is not an easy task. Metadata provide a better understanding of what documents mean. The degree of structure present in data, as well as in the coherent and accurate metadata description, has strong influences on techniques used for search and retrieval. Search techniques essentially rank data relevance and calculate similarities among data. The enormous variety of data and formats used by BER programs, along with poorly structured metadata, outstrips the technical capacity of even the best search algorithms.
2. **A standard model is needed for provenance, curation, and metadata annotation.** Data provenance ensures that updates and corrections to data collections are made in a transparent and traceable way to establish when and why datasets might be altered and corrected. Data do not exist in a vacuum, especially scientific data such as the variety produced in BER programs. To use, interpret, and trust the data, contextual information must be provided on how the data are generated, captured, processed, analyzed, and validated. We lack a comprehensive and standardized set of models that can cope with volume, velocity, variety, and veracity, the unique characteristics of big data, with respect to data provenance, curation, and metadata annotation.
3. **Many scientific challenges require multidomain databases.** Databases are needed for effective running and management of data, particularly for facilitating searching and new data updates. Different forms of data have been generated from various instruments deployed by BER research programs. The natural systems covered by the BER programs are not only structurally and spatially complex with many different interacting parts spanning molecular to global scales, but they also are dynamically complex, encompassing processes that occur over timescales ranging from nanoseconds to centuries. Multidomain databases are desirable to deal with heterogeneous data sources and manage relationship among data.
4. **Cross-facility coordination is needed with data management, data sharing, repositories, metadata, digital object identifiers (DOIs), data formats, data visualizations, curation, provenance, and search.** Global efforts are needed to oversee data management and usage from all perspectives and to establish standards for data and metadata across platforms and research disciplines. Metadata and laboratory methods must be clearly documented and available for any publicly deposited experimental data. Without such documentation, consistency within and across laboratories will

not be attainable, thus creating insurmountable barricades in integrating multiscalar data across systems.

5. **Limited support needs augmentation to provide derived data on demand in a reproducible way, including all metadata, provenance, and derived data products.** In addition to raw data directly from instruments, researchers often combine datasets from various sources (including, but not limited to, facilities), or use computational techniques, thus generating new “derived” datasets. There is currently limited support to provide such derived datasets in a flexible and sustainable way. For example, in a computational pipeline that transforms raw data in a specific way with specific algorithmic parameters, recovering intermediate steps, or the “provenance” of the end result, often proves useful to make further progress without needing to start from the initial data. This is a real challenge when developing and experimenting in a hardware and software marketplace that is moving very rapidly from month to month, or year to year. Additionally, not every dataset leads to a positive result, so the information may not be in a dataset that is published in the literature—still, these data are worth preserving for possible inclusion in future studies.
6. **Single-processor performance has not improved significantly since 2005, yet scientists continue to demand more computing capability for increasingly complex applications.** The computing technology challenges, which will only increase as we near the end of transistor-density scaling (Moore’s Law) within the next decade, mean that increases in performance will come from accelerators (e.g., GPUs), software management memory, manycore architectures, and other features that likely change the way software is written. At the same time, the simulation codes continue to add more physical models, operate across scales, and exploit sparsity and adaptivity in a variety of ways. These application trends tend to be at odds with the needs for large amounts of fine-grained parallelism and high ratios of computation to data movement, all required to take advantage of the hardware. Computing facilities need to balance the demand for more computing with the desire to avoid disrupting software.
7. **Analytics workloads, adding more diversity to existing breadth of simulation codes, making satisfying all user requirements more difficult.** The growing interest in deep-learning algorithms is a good example of how large numbers of GPUs on a single node can be effective for training deep neural nets, but this is not aligned with applications that require larger scale and perform better on more traditional processor architectures.
8. **Integration of observational data into simulations complicates workflows and may require real-time job scheduling, which requires extra capacity to meet surge requirements.** Data analytics jobs may need to process data in real time, including running simulations to solve inverse problems or reconstructing 3D models from images. These needs are at odds with traditional batch-scheduling strategies used in HPC facilities, perhaps leading to inefficient use of computing that is dedicated to a particular project. Commercial clouds offer elastic resources, but large-scale parallel computations are not well suited, and the costs can be high especially for problems that need to move large data volumes.

CHARGE 3 RESPONSE

Development of Additional User Facility Capabilities

In addition to JGI, EMSL and ARM, BER has several projects that run data services, including ESG for climate data, ESS-Dive for Earth science data, MG-RAST for metagenome data, and KBase for a variety of omic data. These projects own and manage their own compute and storage hardware, databases, web servers, or custom software. BER researchers also rely on data services such as NCBI's SRA for sequence read data and PRIDE for proteomic data. These systems and services provide a patchwork of capabilities that are not coordinated, complete, or fully integrated. Building on one of the recommendations in section 1.2, which calls for an infrastructure strategy, in this section we lay out a vision of BER's leadership role in establishing a federated set of national User Facilities for omics-climate-environment computing and data, and some specific gaps that BER should fill.

Federated Data and Computing Infrastructure

This infrastructure will support modeling, simulation, analytics, and learning across data and computational scales. It will provide the computing necessary to analyze next-generation experimental devices, along with on-site computing and storage when needed and streaming to centralized resources when possible. Data in those facilities will be processed for quality, annotated, searched, served to the community, and analyzed, probably multiple times. A federated facility will provide a single entry point for researchers who want to search across datasets and a consistent framework to search and connect across datasets. It will establish community data standards and formats, the sharing and preserving of data, and development of robust workflows that enable reproducibility of research results and simulations. It will support advanced high-end simulations, analytics, and integrated workflows.

While current BER and ASCR facilities will be part of this federated model, there are facility gaps to be filled, namely access to midrange commodity computing and advanced and persistent data services.

Midrange Computing

The need for general-purpose midrange computing leads us to recommend that BER consider procuring (most likely as one or more add-ons to existing facilities) resources dedicated to BER applications. This resource would not replace the high-end cycles provided by NERSC and the LCFs, but instead would supplement these cycles for production-ready codes running on moderate node counts. It also would provide a level of guaranteed access that would help address the allocation decision uncertainty at LCFs. The facility would likely purchase general-purpose processors to support as many codes as possible, including many cutting-edge codes for simulation campaigns that do not require exascale resources. The utilization policies should be adjusted to maximize scientific output even at the expense of obtaining high utilization rates. In particular, the facility should support development activities and other types of simulation campaigns that require fast turnaround and/or real-time computing. These estimates would support the code-development process, where it is critical to obtain useful results on a daily basis. It would also support other types of human-in-the-loop computing, where experts are involved in the active management and control of long-running simulations. A portion of the machine can also be devoted to real-time computing needs, such as is done currently on dedicated resources like those procured by the ARM facility.

Data Preservation and Curation

Preservation and curation of resulting data, in both raw and derived forms, are vital for data discoverability and reusability, evaluation of data and model uncertainty, and data tracking (e.g., DOIs), which will demonstrate the scientific impact of data generated by DOE BER research. BER needs a highly reliable and available facility to store and serve its data, making the data easily discoverable by scientists with varied expertise and interests. Investigators may well understand specific data sources at a facility they use frequently, but, when looking for innovative ways to connect data, finding new data sources that can be brought to bear can provide serendipitous discovery. There is a need to provide a unified framework for querying and identifying available data resources across facilities, not just within one or the other. Organizing metadata from individual projects into a queryable form will be an important component of this system. The data should be centralized in way that all of it can be viewed, search, selected, and downloaded based on meaningful scientific queries. Computing and analytics tools should be co-located with data services, including mining, advanced and multimodal analytics, as well as learning. This will lead to a desire for centralization, while the existing distribution of experimental facilities, observational sites, and data housed by other agencies or programs will lead to some level of distribution. The data facility will almost certainly need to handle some types of protected data and to secure workspaces to handle personally identifiable information, data from industry, or other proprietary information. The proposed system also should be able to automatically query and access data from existing (and widely used) community resources, so that data cuts or slices can be generated on the fly based on user queries. Again, indexing the project metadata available in these data repositories to enable efficient querying will be an important part of this effort.

Recommendations

6.8 Establish a federated data and computing infrastructure.

6.9 Procure dedicated midrange computing.

6.10 Establish data preservation facilities.

CHARGE 4 RESPONSE

Opportunities for Collaboration Among User Facilities

There are multiple opportunities to collaborate across the BER science community, across the Office of Science, and across agencies.

Joint Meetings

Workshops within BER agencies and User Facilities as well as across other agencies and facilities could very well demonstrate the value of interdisciplinary research groups linking informaticists, analysts, and statisticians with biologists. The goal of these workshops will be to provide overviews of proper experimental designs and techniques and analytical methods such as controlling for the quality of data, data analysis, data mining, machine learning, and meaningful interpretation of analytical results. These workshops would show that cross-disciplinary integration within research groups can yield better-planned projects based on obtaining more robust experimental designs, thus saving costs, time, and materials, as well as producing better-powered and statistically sound experiments. However, while workshops are useful in their moments of glory, if methods and techniques presented are not instated or implemented regularly, their efficacy is greatly decreased. Additionally, communication between workshop presenter and workshop attendees is often nonexistent after the workshop, leaving workshop attendees without resources and support.

Integration of informaticists such as bioinformaticians and biostatisticians within a biology research laboratory has many benefits. The intimate hand-in-hand research collaboration between analyst and biologist yields immediate biofeedback between the bio- and the informatics within each project, and may lead to quicker scientific discoveries. In this setting, the analyst has ownership to the research, rather than being a service provider with little investment. Furthermore, the knowledge of the informatics and statistical or mathematical techniques stays and grows throughout the group, thereby breeding a new set of analysts. This seems to be a good solution to the analytical bottleneck, as it increases the number of coowners and co-analysts within specific research projects and thereby reduces the strain on a smaller number of expert analysts acting as service providers in core facilities.

Implementing an idea such as this within the User Facility level should be possible as well. Not only could research groups within each User Facility cross-hybridize to include informaticists, but a cadre of domain-specific mathematics, computer science, and statistics experts could be made available to help users within and across User Facilities as well as within and across agencies.

Coordination Between Agencies

Section 6 of the Grand Challenges Report[31] indicates that intra- and interagency collaborations should be used to leverage efforts of ontology development, data management, and data integration to facilitate data exchange and comparisons across different systems. For example, the National Microbiome Initiative, launched in 2016, is a collaborative effort among the White House Office of Science and Technology Policy (OSTP), several federal agencies, and private-sector stakeholders to support the study of microbiomes across different ecosystems.[32] Similarly, much can be learned from the KBase initiative for biology; this system, although focused on sequence data, may be leveraged at some point to climate and environmental systems, as well as enable easier data exchange. The integration of specific databases developed at other federal agencies (e.g., the Sequence Read Archive at the National Institutes of Health's [NIH] National Center for Biotechnology Information [NCBI], or the NSF's Biological Databases Initiative) would greatly facilitate integration and comparisons of multiscalar and cross-systems data. Current effort and resources could be reduced notably by more deliberate efforts to *coordinate and collaborate among agencies*. To conquer the challenge of public data collection ranging from the Earth sciences to microbiome levels, data and tool exchange must be made continuous and seamless across researchers, irrespective of association to federal agency or other funding source. In other words, an agency-agnostic knowledge discovery effort should be implemented.

Coordination Groups Across BER

Intraagency coordination groups, likely domain specific, enable sharing of best practices and coordinated development of capabilities. Community-driven data coordination groups when resourced to develop joint projects have the potential to produce significant savings and increase sharing of development of emerging cross-facility/project needs including data visualization and analysis tools, data archiving and curation approaches, and model development. These coordination groups can be scheduled to meet concurrent with PI and other meetings that typically would include many of our workshop's participants. These meetings can bring together users or developers to discuss methods and techniques such as for data sharing, maintenance, organization, curation, and annotation.

Coordination Across the Office of Science

An underlying premise of the data and computing challenges is the need for BER to coordinate across the Office of Science, especially with ASCR and BES. This includes long-standing efforts such the ASCR-BER Scientific Discovery through Advanced Computing (SciDAC) program (in partnership with the National

---

[31] BERAC. 2017. *Grand Challenges for Biological and Environmental Research: Progress and Future Vision; A Report from the Biological and Environmental Research Advisory Committee*, DOE/SC-0190, BERAC Subcommittee on Grand Research Challenges for Biological and Environmental Research (science.energy.gov/~/media/ber/berac/pdf/Reports/ BERAC-2017-Grand-Challenges-Report.pdf).

[32] OSTP. 2016. "White House Unveils National Microbiome Initiative," *Nature Biotechnology* **34**(6), 580. DOI: 10.1038/nbt0616-580a.

Nuclear Security Administration [NNSA]), component of the Office of Science–NNSA Exascale Computing Project (ECP), and the use of facilities across the Office of Science, such as the joint BER-ASCR IDEAS project (i.e., Interoperable Design of Extreme-scale Application Software; https://ideas-productivity.org). IDEAS began in 2014 to address issues of software productivity and sustainability in the Office of Science computational science and engineering community, with a particular emphasis on use cases in Subsurface and Terrestrial Ecosystem Modeling. Now supported by ASCR under the Exascale Computing Program (ECP) IDEAS has focus areas that could support beneficial collaboration, including: (1) working with individual application development and software technology teams to understand productivity bottlenecks and improve software development practices through the Productivity and Sustainability Improvement Plan (PSIP) methodology; (2) resources provided through the Better Scientific Software portal (BSSw.io), a community-driven hub for sharing information on practices, techniques, experiences, tools, and other resources to improve developer productivity and software sustainability; and (3) training and tutorial events such as the Best Practices for HPC Software Developers webinar series.

Also representing another example of collaboration within BER are the “Cyberinfrastructure Working Groups (CWG),” established by the Environmental Systems Sciences (ESS) program (joint between the Terrestrial Ecosystem Science [TES] and Subsurface Biogeochemical Research [SBR] programs) to address community needs in model-data integration. The ESS community Cyberinfrastructure aims to enable world-class science by providing capabilities for data ingestion, management and curation, data analysis and visualization, coupled modeling and data publication. These capabilities should exist in a collaborative research environment that allows for *sharing data and results*. The CWG Executive Committee consists of representatives from each of the major ESS projects (e.g., SFAs, NGEEs, and others), but it does not currently include facility representatives. The Executive Committee establishes working groups to address specific community needs such as data or metadata standards and software interface definitions. An opportunity for collaboration exists through participation of BER User Facility representatives in the Executive Committee and Working Groups.

Recommendations

6.11 Hold workshops within BER agencies and User Facilities that link informaticists, analysts, and statisticians with biologists.

6.12 Integrate informaticists (such as bioinformaticians and biostatisticians) within a biology research lab within User Facilities.

6.13 Use intra- and interagency collaborations within DOE to leverage ontology development, data management, and data integration to facilitate data exchange and comparisons across different systems.

6.14 Establish intra-agency coordination groups across BER that are domain specific to enable sharing of best practices and coordinated development of capabilities.

6.15 Coordinate efforts across the DOE Office of Science—between ASCR and BES, for example—continuing long-standing efforts such the ASCR/BER SciDAC program, the component of the SC/NNSA Exascale Computing Project, the use of facilities across SC, and the joint BER/ASCR IDEAS project (Interoperable Design of Extreme-scale Application Software; ideas-productivity.org/).

6.16 Work with individual application development and software technology teams across the Office of Science to understand productivity bottlenecks and improve software development practices through the Productivity and Sustainability Improvement Plan (PSIP) methodology.

6.17 Provide more resources across the Office of Science through the Better Scientific Software portal (BSSw.io), a community-driven hub for sharing information on practices, techniques, experiences, tools, and other resources to improve developer productivity and software sustainability.

6.18 Conduct training and tutorial events across the Office of Science, such as the Best Practices for HPC Software Developers webinar series.

Appendix A. Charge Letter to BERAC from the DOE Office of Science

Department of Energy
Office of Science
Washington, DC 20585

November 3, 2017

Dr. Gary Stacey
Endowed Professor of Plant Science
Divisions of Plant Sciences and Biochemistry
271E Christopher S. Bond Life Sciences Center
University of Missouri
Columbia, MO 65211

Dear Dr. Stacey:

The Office of Biological and Environmental Research (BER) science programs continue to be driven by the Department of Energy's (DOE) basic science, energy, and environmental mission needs. BER increasingly uses a complex systems science approach to advance these science missions. This involves studying complex biological and environmental processes that range from molecular to global scales over time horizons of nanoseconds to centuries and beyond. Our goal is to obtain a holistic and predictive understanding of key biological and environmental systems to address DOE's scientific challenges of the future. Maintaining the capabilities to address future scientific challenges includes the periodic evaluation and alignment of User Facilities to changing scientific research programs.

In 2016, the Biological and Environmental Research Advisory Committee (BERAC) was charged to review the 2010 report (DOE/SC-0135), "Grand Challenges for Biological and Environmental Research: A Long-Term Vision", review research that occurred since that report, and develop a new long-term strategic vision. The new report (DOE/SC-0190), completed in 2017, has identified a number of grand challenges that are important to the DOE mission and that utilize BER expertise. A more focused effort is needed to understand and identify the roles of User Facilities to meet the new or revised grand research challenges. For purposes of this exercise, User Facilities should include both national User Facilities and community research infrastructure that enables and allows for community research participation.

I request that BERAC use its combined expertise across the BER portfolio to evaluate the following topics regarding the current and future utilization of User Facilities for BER research:

1. Optimal alignment of User Facilities to support the current BER research portfolio
2. Optimal alignment of User Facilities to support future research needs identified in the 2017 Grand Challenges report (DOE/SC-0190)

Printed with soy ink on recycled paper

3. Development of additional User Facility capabilities
4. Opportunities to collaborate between User Facilities (internal to DOE and also external interagency partners)

I request that BERAC evaluate the aforementioned topics and summarize the findings in a report. I would like to receive the final report by the fall 2018 BERAC meeting. Many thanks for your contributions to this important effort.

Sincerely,

J. Stephen Binkley
Acting Director
Office of Science

cc: Sharlene Weatherwax
Tristram West

Appendix B. List of Recommendations

Biological Systems Science

2.1 Develop metabolic pathway databases based on experimentally annotated gene function and integrate metabolic data needed to achieve subcellular organization of metabolites, enzymes, and pathways.

2.2 Develop structural libraries for metabolites and enzymes.

2.3 Obtain equipment designed to dynamically measure intracellular and interspecies fluxes of metabolism and transport by developing imaging and isotope labeling technologies, applicable to organisms interacting in complex communities *in situ*.

2.4 Develop methods for *in situ* measurements and single-cell measurements.

2.5 Integrate molecular dynamics simulations into flux measurements.

2.6 Develop stoichiometric and kinetic models of metabolism that integrate omic data and allow the transition from observations of changes in gene expression to metabolic activity.

2.7 Establish capacity at JGI or EMSL for stable isotope probing.

2.8 Develop KBase to support quantitative interpretation of isotopomer-enabled metabolic flux modeling of central metabolism and other specific metabolic pathways.

2.9 Train an interdisciplinary workforce for improving the understanding of metabolism.

2.10 Improve methods for large DNA construct design and assembly and expand the availability of rapid prototyping systems.

2.11 Develop and deploy technologies to engineer previously genetically intractable organisms.

2.12 Develop new methods for facile porting of biosynthetic pathways between organisms to investigate the role of physiological context in gene expression and metabolism.

2.13 Expand the tools available for genome-wide genetic disruption and their application to a range of organisms.

2.14 Develop cellular sensors for monitoring metabolism and metabolic state in organisms and how they are influenced by their ecosystem.

2.15 Enhance capability for *de novo* DNA synthesis and assembly of large DNA molecules.

2.16 Build a highly dynamic, shared search platform among JGI, EMSL, and KBase to enable data correction and method sharing outside of static publications.

2.17 Establish and adhere to metadata standards for JGI, EMSL, and KBase and lead efforts in setting such standards in collaboration with other large scientific organizations.

2.18 Leverage ASCR compute resources for data storage and large-scale computing.

2.19 Issue joint funding calls with ASCR to encourage collaboration among biologists, mathematicians, and computer scientists on the development of methods for multimodal data integration and understandable machine learning.

2.20 Develop high-throughput computational methods to better predict function of gene products, including expanding the User Facility computational biology team.

2.21 Develop expression platforms capable of generating sufficient protein for characterization of protein structure and function.

2.22 Employ genome-wide gene disruption and gene expression technologies such as CRISPR, Tn-Seq, and Dub-Seq, which allow for systematic assignment of critical genes under specific conditions.

2.23 Deploy integrated multiomics technologies to understand genome-wide changes in gene expression and metabolism.

2.24 Enhance the integration of “omic” and other data generated at multiple User Facilities by enhancing the coordination among these facilities.

2.25 Develop facilities to better characterize phenotypes resulting from altered gene function, including whole-organism and population growth and development, as well as high-resolution imaging and monitoring of metabolic changes.

2.26 Expand the accessibility of microfluidic and nanotechnology techniques for high-throughput, *in situ*, deep phenotyping and single-cell applications.

2.27 Develop facilities so that researchers can perform imaging on a sample and then subject these samples to omics approaches.

2.28 Invest in state-of-the-art high-throughput cryo-EM instrumentation and couple with cell-free expression capabilities within the BER network to facilitate rapid structure determination and protein annotation.

2.29 Deploy new cryo-ET capabilities within already established User Facilities for multimodal interrogation of whole cells and ease of access by the broad user community.

2.30 Extend the portfolio of microscopic imaging facilities designed to perform label-free imaging available to BER users.

2.31 Establish a Coordinated Network for Systems Biology, a multisite network comprising existing BER and other DOE User Facilities that coordinates multiomics approaches performed with broad spatial and temporal scales to address large-scale and complex challenges for understanding biological systems

Earth and Environmental Systems

3.1 Develop new technologies that address persistent scientific needs for ARM, including convective vertical velocity, aerosol profiles, ice nucleation, and continuous thermodynamic profiling. Technologies warranting investment to meet these needs include unmanned aircraft systems and tethered balloon instrumentation and miniaturization to access previously inaccessible domains.

3.2 Employ targeted calls for User Facilities to better address specific Grand Challenges with a focus on cross-disciplinary and coupled system studies.

3.3 Consider the mechanisms used by other user communities (e.g., astronomers and high-energy experimental particle physicists) to evaluate and select from candidate augmentations to existing User Facilities. The mechanisms should start from predefined evaluation metrics and definitions of success and should operate in an open and transparent manner with extensive documentation of the prioritization procedures.

3.4 Employ targeted calls for the design of several new User Facilities to better address specific Grand Challenges and emerging research frontiers.

3.5 Identify and determine for each User Facility the controlling emergent processes and behaviors of a system (e.g., the “rare biosphere”) at different scales as a means to better constrain how these processes interact across scales and to prioritize facility activities.

3.6 Augment and align ARM resources along a central U.S. transect to better address both small- and large-scale processes associated with the full life cycle of convective precipitation across the central United States. Such a transect could include new sites in Colorado, SGP, and the southeastern United States, with smaller sites in between, links to other networks, and integration with other key environmental transitions (e.g., forest coverage, drought, ecosystem processes, and carbon cycle).

3.7 Develop and employ an appropriate cross-scale modeling framework for each primary User Facility as an instrument to support up- and down-scaling between observations and large-scale models (e.g., LASSO for ARM).

3.8 Employ advanced unmanned aircraft systems in a systematic approach to bridge across scales and to assess the spatial representativity of User Facility observations in a variety of multiscale environments.

3.9 Develop a network of AmeriFlux omics-to-ecosystems supersites, where high–temporal resolution field and laboratory observations of omics, microhabitat-scale conditions, and fluctuating resources are generated automatically and data are compared with ecosystem flux observations and models.

3.10 Build new capacity through a combination of AmeriFlux and ARM technologies to map individual tree structure and seasonal and interannual forest dynamics across the network.

3.11 Establish a joint facility activity among EMSL, JGI, and ARM, perhaps by extending existing Facilities Integrating Collaborations for User Science (FICUS) collaborations, to develop and implement a comprehensive observational strategy (field and laboratory) to measure and discern modes of ice nucleation under real atmospheric conditions.

3.12 Develop a cloud chamber with the ability to examine aerosol particle formation and cloud activity, with links to EMSL for characterization of organic INPs formed through (photo)-chemical processing of organic precursor emissions.

3.13 Deploy the ARM Mobile Facility No. 3 for extended operations at a location relevant for addressing cryosphere impacts on sea level, such as West Antarctica or Southern Greenland.

3.14 Hold a targeted workshop to explicitly consider how BER facilities can address cryospheric change.

3.15 Hold a targeted workshop that builds on the prior Terrestrial-Aquatic Interface workshop, broadening the scope to include areas further from the coastlines in both directions as well as the impacts of the changing cryosphere. Workshop outcome: Framework for how User Facilities can address evolving needs on this theme.

3.16 Develop a framework that leverages the full suite of capabilities at EMSL and JGI to conduct manipulative experiments using ecotrons.

3.17 Consider a targeted research announcement aimed at supporting manipulative field experiments that leverage EMSL and JGI collaborations.

3.18 Establish a User Facility to enable manipulative experiments at field-relevant scales that are critical for advancing our understanding of the linkages between physical and biological systems and across scales of organization, from molecules to habitats to ecosystems.

3.19 Encourage concerted coordination between DOE ASCR and the BER User Facilities to improve the pace of data archival, the quality of metadata, the ease of data access, and tools for data analysis.

3.20 Develop a living and broadly accessible repository of analysis tools, with collaborative links to the various research programs that support the tools' development and use.

3.21 Consider aggregating tools for data analysis, data-model synthesis, and state-of-the-art simulation modeling into a software container that could be used at users' institutions, on User Facility computational resources, or in the cloud. This would help maximize the range of options and efficiency for analysis of User Facility data and of community models.

3.22 Implement the call in the Grand Challenges report to develop a computational and synthesis User Facility that supports the rapid design, generation, evaluation, and diagnosis of ESMs, including robust data-model synthesis. This facility will support the accessibility and availability of models and simulations to a wide community of potential users; and the development of new models addressing scaling across organization over the full purview of BER (omics to Earth).

3.23 Encourage joint focus on Grand Challenge–relevant scientific themes through (1) coordinated User Facility activities, where linkages are well established, and (2) workshops to develop a vision for coordinated efforts to address cross-disciplinary themes in the Grand Challenges.

3.24 Establish stronger links between the operational model and satellite communities to explore more effective transfer of knowledge between BER facilities and these platforms via assimilation, assessment, and intercomparison.

3.25 Further develop and implement a framework for joint calls, review, and decision making (perhaps via the FICUS program): (1) across multiple User Facilities to enable and incentivize cross-disciplinary research to address joint research priorities and Grand Challenges and (2) across User Facilities and appropriate science programs to ensure the availability and effective use of scientific resources. The primary focus for such a framework may be internal to BER, but it should also consider engagement from external agencies and facilities. Such joint calls could be supported through dedicated crosscutting budgets for integrative research.

3.26 Strengthen the connection of "capacity building" programs with specific User Facilities and specific Grand Challenge themes.

3.27 Consider cross–User Facility summer schools or advanced training activities that bring together diverse groups of students and scientists, organized around leveraging User Facility capabilities for specific Grand Challenge themes.

Microbial to Earth System Pathways

4.1 Expand EMSL computational support staff and their expertise to include the array of applications and codes relevant to BER users.

4.2 Institute a time delay before data are released, until publication, or for one year after a user project ends, whichever comes first. For projects with components at different User Facilities, match the time frames of the project components as well as the time delays for data release.

4.3 Shift weight toward metrics of User Facility success that recognize facility efforts in maintaining a productive, returning user base, rather than weighting toward total numbers of users served.

4.4 Enable process modeling and data-related computation by investing in midrange computing infrastructure and personnel time.

4.5 Develop a robust computational framework that can connect and inform models at multiple scales and that facilitates iteration based on input from experimental and field data and modeling output.

4.6 Develop field deployable, multimodal, remotely controlled sensors that ideally conduct nondestructive measurements to (1) characterize how microbial habitat–scale heterogeneity and dynamics influence biogeochemical processes and (2) validate relevance of lab experiments in field.

Energy and Environmental Resilience

5.1 Establish a strategically distributed network of research centers focusing on the Science of Energy and Environmental Resilience (SEER) that would develop and apply an array of capabilities for evaluating and projecting the dynamics of coupled human and environmental systems in support of national needs.

Computation and Data Analysis

6.1 Provide tools at facilities for labeling, metadata management, and data discovery both within one facility and across DOE and non-DOE facilities.

6.2 Provide tools at facilities to manage derived data products, as well as long-term storage of raw data whenever possible.

6.3 Develop an infrastructure strategy that addresses data analysis and storage needs.

6.4 Analyze the most important applications and determine which ones will need significant performance increases to meet scientific demands and which can continue with only modest increases.

6.5 Work with ASCR to ensure continued access to testbeds with emerging architectures, as well as to training and user programs to help with the evaluation and code transitioning of critical BER applications, among others.

6.6 Work with the research community and computational facilities to determine the hardware, software, and usage policies needed to support researchers' complex workflows.

6.7 Address the needs of real-time streaming data and interactive computing as part of the recommended infrastructure strategy.

6.8 Establish a federated data and computing infrastructure.

6.9 Procure dedicated midrange computing.

6.10 Establish data preservation facilities.

6.11 Hold workshops within BER agencies and User Facilities that link informaticists, analysts, and statisticians with biologists.

6.12 Integrate informaticists (such as bioinformaticians and biostatisticians) within a biology research lab within User Facilities.

6.13 Use intra- and interagency collaborations within DOE to leverage ontology development, data management, and data integration to facilitate data exchange and comparisons across different systems.

6.14 Establish intra-agency coordination groups across BER that are domain specific to enable sharing of best practices and coordinated development of capabilities.

6.15 Coordinate efforts across the DOE Office of Science—between ASCR and BES, for example—continuing long-standing efforts such the ASCR/BER SciDAC program, the component of the SC/NNSA Exascale Computing Project, the use of facilities across SC, and the joint BER/ASCR IDEAS project (Interoperable Design of Extreme-scale Application Software; ideas-productivity.org).

6.16 Work with individual application development and software technology teams across the Office of Science to understand productivity bottlenecks and improve software development practices through the Productivity and Sustainability Improvement Plan (PSIP) methodology.

6.17 Provide more resources across the Office of Science through the Better Scientific Software portal (BSSw.io), a community-driven hub for sharing information on practices, techniques, experiences, tools, and other resources to improve developer productivity and software sustainability.

6.18 Conduct training and tutorial events across the Office of Science, such as the Best Practices for HPC Software Developers webinar series.

Appendix C. BERAC Subcommittee Workshop Agenda

Biological and Environmental Research Advisory Committee (BERAC)
Subcommittee on User Research Facilities
Working Meeting
Hilton Gaithersburg
April 23-24, 2018

Agenda

Monday, 23 April 2018

8:00	Meet, coffee
8:30 – 8:45	Introduction: clarify goals, pathway to get there, grand challenges, where we are in answering the charge [Bruce Hungate]
8:45 – 10:00	Brief presentations with Q&A from ARM, JGI, and EMSL
10:15 – 10:30	Issue the charge (BH)
10:30 – 12:00	Working groups discussion I ○ Generative; add to outlines; structure by focusing on recommendations
12:00 – 13:30	Lunch and outside
13:30 – 15:30	Working groups discussion II ○ Refine recommendations; prioritize recommendations (don't delete any ideas yet) ○ Post Drafts to Google Drive to be available for reference in cross pollination session
15:30 – 15:45	Break
15:45 – 17:30	Cross pollination ○ Mix up working groups ○ Each new group evaluates one other working group's draft outline – what's missing? – are priorities appropriate? – what ideas occur in multiple groups?
17:30 – 18:00	Working groups reconvene and refine outline for summary
17:15 – 20:00	Reconvene as large group for Summary, Working Dinner ○ Each group presents ○ Where are we? ○ What are emerging themes? ○ What more do we need?

Tuesday, 24 April 2018

8:00	Meet, coffee
8:30	Charge, goals, where we are [Bruce Hungate]
8:30 – 11:30	Working Groups Write, google docs saved
11:30 – 12:00	Reconvene as large group for Summary

Appendix D. BERAC Members and Workshop Participants

Biological and Environmental Research Advisory Committee

Gary Stacey, Chair
University of Missouri

Bruce A. Hungate, Vice Chair
Northern Arizona University

Sarah M. Assman
Pennsylvania State University

Julie S. Biteen
University of Michigan

Amy M. Brunner
Virginia Tech

James J. Hack
Oak Ridge National Laboratory

Andrzej Joachimiak
Argonne National Laboratory

Kerstin Kleese van Dam
Brookhaven National Laboratory

Cheryl R. Kuske
Los Alamos National Laboratory

L. Ruby Leung
Pacific Northwest National Laboratory

Gerald A. Meehl
National Center for Atmospheric Research

Jerry M. Melillo
Marine Biological Laboratory

Gloria K. Muday
Wake Forest University

Kristala L. Jones Prather
Massachusetts Institute of Technology

James T. Randerson
University of California, Irvine

Patrick Reed
Cornell University

G. Philip Robertson
Michigan State University

Karen Schlauch
Desert Research Institute

Daniel Segre
Boston University

Matthew D. Shupe
University of Colorado and NOAA Earth System Research Laboratory

David A. Stahl
University of Washington

John P. Weyant
Stanford University

Huimin Zhao
University of Illinois at Urbana-Champaign

Subcommittee on Scientific User Research Facilities

Bruce A. Hungate, Chair
Northern Arizona University

Zoe Cardon
Marine Biology Laboratory

Serita Frey
University of New Hampshire

Anne Giblin
Marine Biology Laboratory

Mike Goulden
University of California, Irvine

Samantha Joye
University of Georgia

Marcus Kleber
Oregon State University

Gloria Muday
Wake Forest University

Kristala Prather
Massachusetts Institute of Technology

Jim Randerson
University of California, Irvine

Patrick Reed
Cornell University

Karin Remington
Computationality, LLC

Phil Robertson
Michigan State University

Karen Schlauch
Desert Research Institute

Matthew D. Shupe
University of Colorado and NOAA Earth System Research Laboratory

Mark Taylor
Sandia National Laboratory

John Weyant
Stanford University

Kathy Yelick
University of California, Berkeley, and Lawrence Berkeley National Laboratory

Working Group Participants

Deb Agarwal
Lawrence Berkeley National Laboratory

Amy Callaghan
University of Oklahoma

Katherine Calvin
Pacific Northwest National Laboratory

Zoe Cardon
Marine Biology Laboratory

William Collins
Lawrence Berkeley National Laboratory

Wei Ding
University of Maryland

James Evans
Pacific Northwest National Laboratory

Kjiersten Fagnan
DOE Joint Genome Institute, Lawrence Berkeley National Laboratory

Karen Fischer-Vanden
Pennsylvania State University

Serita Frey
University of New Hampshire

Anne Giblin
Marine Biology Laboratory

Mike Goulden
University of California, Urvine

Steve Hamilton
Michigan State University

Mohamad Hejazi
Pacific Northwest National Laboratory

Kirsten Hofmockel
Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory

Bruce A. Hungate
Northern Arizona University

Michael Jensen
Brookhaven National Laboratory

Samantha Joye
University of Georgia

Marcus Kleber
Oregon State University

Ian Kraucunas
Pacific Northwest National Laboratory

Lee Ann McCue
Pacific Northwest National Laboratory

Sean McSweeney,
National Synchrotron Light Source-II, Brookhaven National Laboratory

Gloria Muday
Wake Forest University

Kristala Prather
Massachusetts Institute of Technology

Jim Randerson
University of California, Irvine

Patrick Reed
Cornell University

Karin Remington
Computationality, LLC

Phil Robertson
Michigan State University

Tim Scheibe
Pacific Northwest National Laboratory

Karen Schlauch
Desert Research Institute

Matthew Shupe
University of Colorado and NOAA Earth System Research Laboratory

Ginash Siram
University of Maryland

Mark Taylor
Sandia National Laboratories

John Weyant
Stanford University

Kathy Yelick
Lawrence Berkeley National Laboratory and University of California, Berkeley,

Shibu Yooseph
University of Central Florida

User Facility Community Representation

Adam Arkin
KBase, Lawrence Berkeley National Laboratory

Larry Berg
Atmospheric Radiation Measurement (ARM), User Executive Committee

Charles Black
Center for Functional Nanomaterials

Harvey Bolton
Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory

Ben Brown (responding for Barbara Helland)
Advanced Scientific Computing Research, DOE

Hans Christen
Center for Nanophase Materials Science, Oak Ridge National Laboratory

Sharon Collinge
National Ecological Observatory Network

Ed DeLong
Joint Genome Institute

Kelly Gafney
Stanford Synchrotron Radiation Lightsource, SLAC

Supratik Guha
Center for Nanoscale Materials Science, Argonne National Laboratory

James Hack
Oak Ridge Leadership Computing Facility, Oak Ridge National Laboratory

Paul Hanson
SPRUCE, Oak Ridge National Laboratory

Britt Hedman
SLAC National Accelerator Laboratory, Stanford University

John Hill
National Synchrotron Light Source II, Brookhaven National Laboratory

Keith Hodgson
SLAC National Accelerator Laboratory, Stanford University

Steve Kevan
Advanced Light Source, Lawrence Berkeley National Laboratory

Paul Langan
High Flux Isotope Reactor, Oak Ridge National Laboratory

Jim Mather
Atmospheric Radiation Measurement (ARM), Pacific Northwest National Laboratory

Sean McSweeney
National Synchrotron Light Source-II, Brookhaven National Laboratory

Nigel Mouncey
DOE JGI, Lawrence Berkeley National Laboratory

Jeff Neaton
The Molecular Foundry

Andreas Roelofs
Center for Integrated Nanotechnologies

Lou Sherman
Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory

Stephen Streiffer
Advanced Photon Source, Argonne National Laboratory

Alan Tennant
High Flux Isotope Reactor, Oak Ridge National Laboratory

Margaret Torn
AmeriFlux Management Project, Lawrence Berkeley National Laboratory

Soichi Wakatsuki
SLAC National Accelerator Laboratory, Stanford University

Stan Wullschleger
Next Generation Ecosystem Experiment–Arctic, Oak Ridge National Laboratory

Appendix E. Acronyms and Abbreviations

2D, 3D	two-, three-dimensional
AGU	American Geophysical Union
ALCF	Argonne Leadership Computing Facility
ALS	Advanced Light Source (at LBNL)
AMD	advanced micro device
ARM	Atmospheric Radiation Measurement User Facility
ARMBE	ARM Best Estimate
ASCR	Office of Advanced Scientific Computing Research (DOE Office of Science)
ASLO	Association for the Sciences of Limnology and Oceanography
BECCS	bioenergy with carbon capture and storage
BER	Office of Biological and Environmental Research (DOE Office of Science)
BERAC	Biological and Environmental Research Advisory Committee
BES	Office of Basic Energy Sciences (DOE Office of Science)
BNL	Brookhaven National Laboratory
BRC	Bioenergy Research Center (DOE)
BSS	Biological Systems Science
BSSD	Biological Systems Science Division (BER)
CAMERA	Center for Applied Mathematics for Energy Research Application (ASCR and BES)
CAPT	Cloud-Associated Parameterizations Testbed (ARM)
CDA	Computation and Data Analysis
CESD	Climate and Environmental Sciences Division (BER)
CESM	Community Earth System Model
CH_4	methane
CIV	Coordination, Integration, and Visualization (SEER)
CNSB	Coordinated Network for Systems Biology (proposed)
CO_2	carbon dioxide
CPU	central processing unit
CRAGE	Chassis-Independent Recombination Assisted Genome Engineering (universal strain-engineering platform)
CRISPR	clustered regularly interspaced short palindromic repeats (gene-editing technology)
cryo-EM	cryogenic electron microscopy
cryo-ET	cryo-electron tomography
cryo-SXT	cryogenic soft X-ray nanotomography
CWG	Cyberinfrastructure Working Groups (BER's TES and SBR)
DMZ	special local network configuration designed to improve security via a firewall
DOE	U.S. Department of Energy
DOI	digital object identifier
E3SM	Energy Exascale Earth System Model (DOE)
EBI	European Bioinformatics Institute (EMBL)
EcoFAB	Fabricated Ecosystems (cross-functional team at LBNL)
ECP	Exascale Computing Project (ASCR)

ECR	Early Career Research
EER	Energy and Environmental Resilience (BER)
EES	Earth and Environmental Systems (BER)
EFRC	Energy Frontiers Research Center (BES)
ELM	Export Land Model
EM	electron microscopy
EMBL	European Molecular Biology Laboratory
EMSL	Environmental Molecular Sciences Laboratory (DOE, at PNNL)
EPR	electron paramagnetic resonance
ESA	Ecological Society of America
ESG	Earth Systems Grid (DOE)
ESGF	Earth System Grid Federation (international effort)
ESM	Earth System Model
ESnet	Energy Sciences Network (DOE)
ESR	Energy Sustainability and Resilience (BER)
ESS	Environmental Systems Sciences (BER'S TES and SBR)
ETOP	Emerging Technologies Opportunity Program (JGI)
FEBA	Functional Encyclopedia of Bacteria and Archaea (LBNL)
FICUS	Facilities Integrating Collaborations for User Science (BER)
FOA	Funding Opportunity Announcement
FPGA	field-programmable gate array
FTICR	Fourier transform ion cyclotron resonance
GCAM	global change assessment model
GPU	graphics processing unit
GTAP	Global Trade Analysis Project (international, coordinated at Purdue University)
HPC	high-performance computing
IDEAS	Interoperable Design of Extreme-scale Application Software (BER and ASCR)
iHESD	Integrating Human and Earth System Dynamics (SFA at PNNL)
ILAMB	International Land Model Benchmarking project
IM3	Integrated Multisector, Multiscale Modeling
IMG/M	Integrated Microbial Genomes and Microbiomes (JGI)
IPCC	Intergovernmental Panel on Climate Change
IR-SNOM	infrared scanning near-field optical microscopy
ITCZ	Inter-Tropical Convergence Zone (NSF)
JGI	Joint Genome Institute (BER)
kb	kilobases
KBase	DOE Systems Biology Knowledgebase (BER)
LASSO	LES ARM Symbiotic Simulation and Observation
LBNL (Berkeley Lab)	Lawrence Berkeley National Laboratory
LCA	life-cycle assessment
LCF	Leadership Computing Facility
LES	large-eddy simulation
LiDAR	Light Detection and Ranging

LTER	Long-Term Ecological Research Network (NSF)
MEND	microbially explicit carbon model
MESP	Microbial to Earth System Pathways (BER)
MS	mass spectrometry
Nano-SIMS	nanoscale-secondary ion mass spectrometry
NASA	National Aeronautics and Space Administration
NCAR	National Center for Atmospheric Research
NCBI	National Center for Biotechnology Information (NIH)
NCEAS	National Center for Ecological Analysis and Synthesis
NEON	National Ecological Observatory Network (NSF)
NERSC	National Energy Research Scientific Computing Center (DOE Office of Science, at LBNL)
NESAP	NERSC Exascale Science Application Program (DOE)
NEST	Network of Energy Sustainability Testbeds
NEXRAD	Next Generation Weather Radar (NOAA)
NGEE	Next-Generation Ecosystem Experiments
NGO	non-governmental organization
NIH	National Institutes of Health
NMR	nuclear magnetic resonance
NNSA	National Nuclear Security Administration (DOE)
NOAA	National Oceanic and Atmospheric Administration
NRC	National Research Council (National Institute of Standards and Technology)
NSF	National Science Foundation
OLCF	Oak Ridge Leadership Computing Facility (ORNL)
ORNL	Oak Ridge National Laboratory
OSSEs	Observing System Sensitivity Experiments (NSF SOCCOM)
OSTI	Office of Scientific and Technical Information (DOE)
OSTP	Office of Science and Technology Policy (White House)
PALM	photoactivated localization microscopy
PB	petabytes
PCHES	Program on Coupled Human and Earth Systems
PCMDI	Program for Climate Model Diagnosis and Intercomparison
PI	principal investigator
PNNL	Pacific Northwest National Laboratory
PRIDE	PRoteomics IDEntifications database (EMBL EBI)
PSIP	Productivity and Sustainability Improvement Plan (scientific software)
PyART	Python ARM Radar Toolkit
SANS	small-angle neutron scattering (microscopy)
SAXS	small-angle X-ray scattering (microscopy)
SBR	Subsurface Biogeochemical Research (BER)
SCGSR	Graduate Student Research program, Office of Science
SciDAC	Scientific Discovery through Advanced Computing (ASCR)
SCM	Single Column Models (ARM)
SEER	Science of Energy and Environmental Resilience (proposed, BER)
SEM	scanning electron microscopy

SESYNC	National Socio-Environmental Synthesis Center (NSF, at University of Maryland)
SFA	Science Focus Area
SGP	Southern Great Plains region
SOCCOM	Southern Ocean Carbon and Climate Observations and Modeling (NSF)
SOM	soil organic material
SP	superparameterized
SPLAT-MS	Single-Particle Laser Ablation Time-of-Flight MS
SPRUCE	Spruce and Peatland Responses Under Changing Environments (BER TES)
SRA	Sequence Read Archive (NIH NCBI)
SSD	Solid State Disk
SRS	stimulated Raman scattering
STORM	stochastic optical reconstruction microscopy
TAI	terrestrial-aquatic interface
TB	terabyte (fourth power of 1,000)
TEM	transmission electron microscopy
TES	Terrestrial Ecosystem Science (BER)
UCAR	University Corporation for Atmospheric Research
VAPs	value-added data products
WDTS	Workforce Development for Teachers and Scientists, DOE