U.S. Department of Energy Office of Science

A Unified Data Infrastructure for Biological and Environmental Research

Report by the BER Advisory Committee

DRAFT:
March 19, 2024 3:51 PM

Biological and Environmental Research Program

U.S. DEPARTMENT OF ENERGY | Office of Science
Biological and Environmental Research

A Unified Data Infrastructure for Biological and Environmental Research

Report from the BER Advisory Committee

Subcommittee Steering Group

Kerstin Kleese van Dam, Chair
Brookhaven National Laboratory

Adam Schlosser, Co-Chair
Massachusetts Institute of Technology

Jeremy Schmutz, Co-Chair
HudsonAlpha Institute for Biotechnology

Ben Bond-Lamberty
Pacific Northwest National Laboratory

Kjiersten Fagnan
Lawrence Berkeley National Laboratory

Ann M. Fridlind
National Aeronautics and Space Administration

Susan Gregurick
National Institutes of Health

Dev Niyogi
University of Texas, Austin

Daniel Segrè
Boston University

Pamela Weisenhorn
Argonne National Laboratory

Designated Federal Officer

Tristram West
U.S. Department of Energy
Biological and Environmental Research Program

About BERAC

The Biological and Environmental Research Advisory Committee (BERAC) provides advice on a continuing basis to the U.S. Department of Energy's (DOE) Office of Science Director on the many complex scientific and technical issues that arise in developing and implementing DOE's Biological and Environmental Research program (science.osti.gov/ber/berac).

Suggested Citation

BERAC. 2024. A Unified Data Infrastructure for Biological and Environmental Research: Report from the BER Advisory Committee, DOE/SC-XXXX. XXXX and XXXX, eds. Biological and Environmental Research Advisory Committee. DOI:XXXXXX.

DOE/SC-XXXX

A Unified Data Infrastructure for Biological and Environmental Research

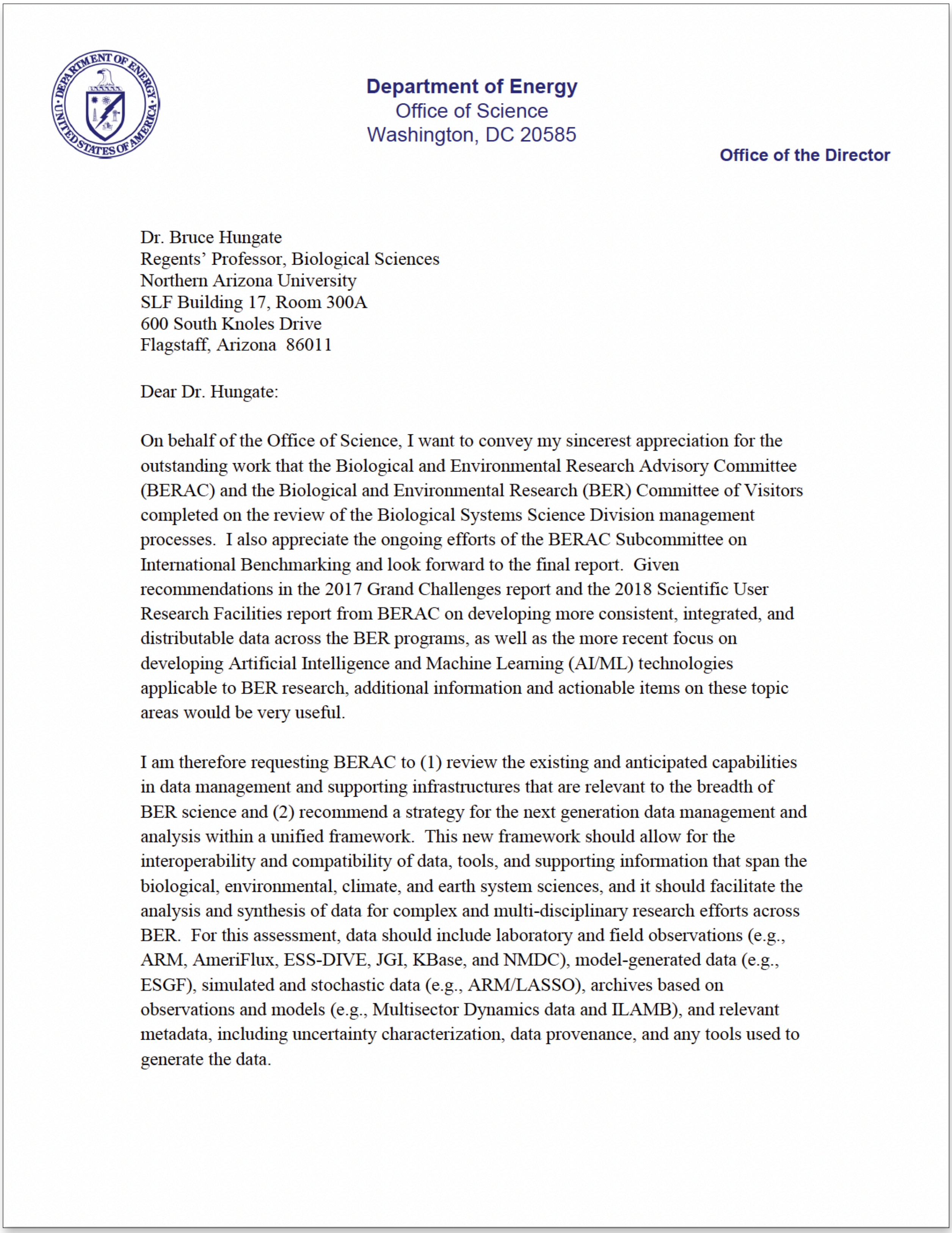
Report from the BER Advisory Committee

March 2024

U.S. DEPARTMENT OF ENERGY | Office of Science

Biological and Environmental Research Program

Charge Letter



Department of Energy
Office of Science
Washington, DC 20585

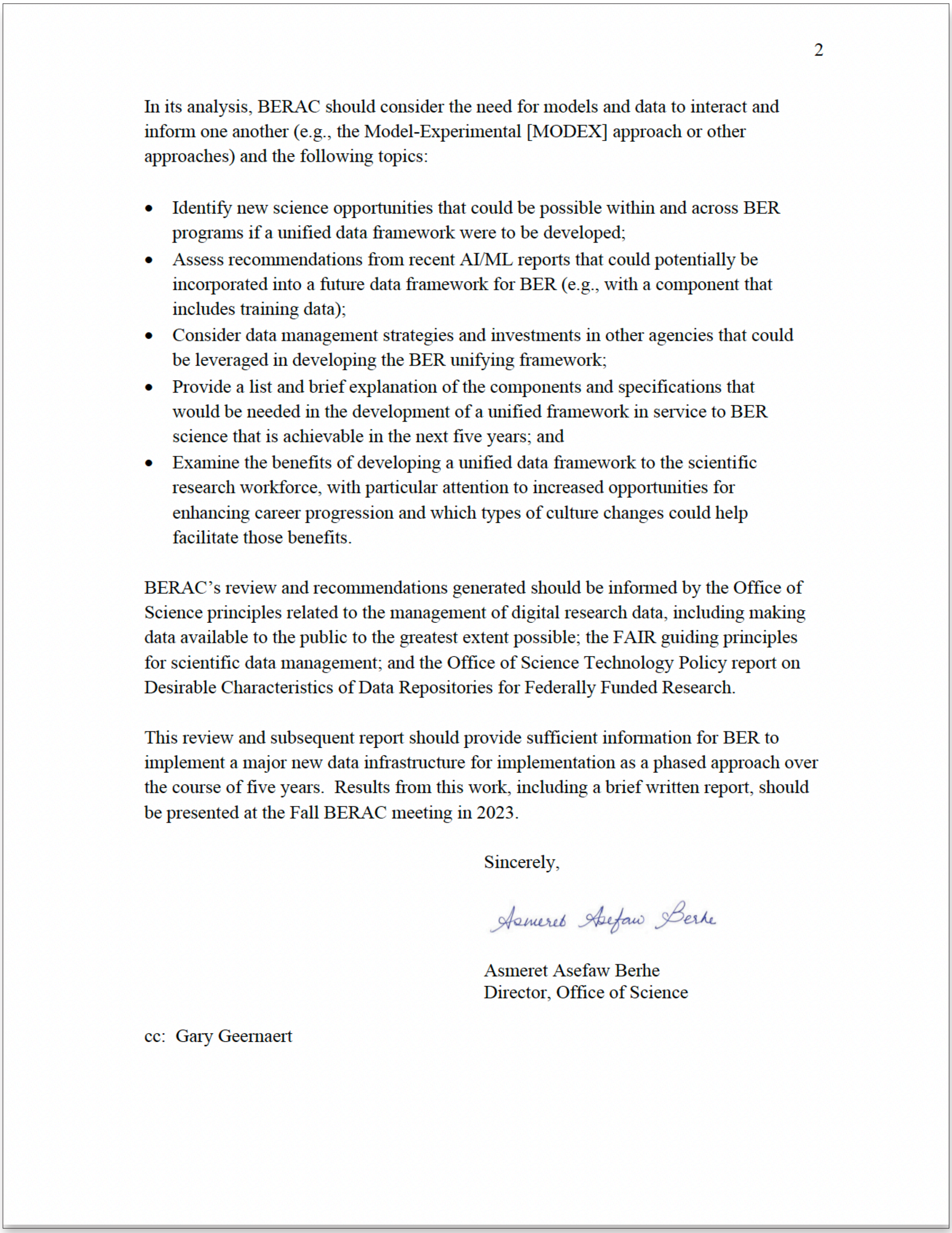
Office of the Director

Dr. Bruce Hungate
Regents' Professor, Biological Sciences
Northern Arizona University
SLF Building 17, Room 300A
600 South Knoles Drive
Flagstaff, Arizona 86011

Dear Dr. Hungate:

On behalf of the Office of Science, I want to convey my sincerest appreciation for the outstanding work that the Biological and Environmental Research Advisory Committee (BERAC) and the Biological and Environmental Research (BER) Committee of Visitors completed on the review of the Biological Systems Science Division management processes. I also appreciate the ongoing efforts of the BERAC Subcommittee on International Benchmarking and look forward to the final report. Given recommendations in the 2017 Grand Challenges report and the 2018 Scientific User Research Facilities report from BERAC on developing more consistent, integrated, and distributable data across the BER programs, as well as the more recent focus on developing Artificial Intelligence and Machine Learning (AI/ML) technologies applicable to BER research, additional information and actionable items on these topic areas would be very useful.

I am therefore requesting BERAC to (1) review the existing and anticipated capabilities in data management and supporting infrastructures that are relevant to the breadth of BER science and (2) recommend a strategy for the next generation data management and analysis within a unified framework. This new framework should allow for the interoperability and compatibility of data, tools, and supporting information that span the biological, environmental, climate, and earth system sciences, and it should facilitate the analysis and synthesis of data for complex and multi-disciplinary research efforts across BER. For this assessment, data should include laboratory and field observations (e.g., ARM, AmeriFlux, ESS-DIVE, JGI, KBase, and NMDC), model-generated data (e.g., ESGF), simulated and stochastic data (e.g., ARM/LASSO), archives based on observations and models (e.g., Multisector Dynamics data and ILAMB), and relevant metadata, including uncertainty characterization, data provenance, and any tools used to generate the data.



In its analysis, BERAC should consider the need for models and data to interact and inform one another (e.g., the Model-Experimental [MODEX] approach or other approaches) and the following topics:

- Identify new science opportunities that could be possible within and across BER programs if a unified data framework were to be developed;
- Assess recommendations from recent AI/ML reports that could potentially be incorporated into a future data framework for BER (e.g., with a component that includes training data);
- Consider data management strategies and investments in other agencies that could be leveraged in developing the BER unifying framework;
- Provide a list and brief explanation of the components and specifications that would be needed in the development of a unified framework in service to BER science that is achievable in the next five years; and
- Examine the benefits of developing a unified data framework to the scientific research workforce, with particular attention to increased opportunities for enhancing career progression and which types of culture changes could help facilitate those benefits.

BERAC's review and recommendations generated should be informed by the Office of Science principles related to the management of digital research data, including making data available to the public to the greatest extent possible; the FAIR guiding principles for scientific data management; and the Office of Science Technology Policy report on Desirable Characteristics of Data Repositories for Federally Funded Research.

This review and subsequent report should provide sufficient information for BER to implement a major new data infrastructure for implementation as a phased approach over the course of five years. Results from this work, including a brief written report, should be presented at the Fall BERAC meeting in 2023.

Sincerely,

Asmeret Asefaw Berhe

Asmeret Asefaw Berhe
Director, Office of Science

cc: Gary Geernaert

BERAC Subcommittee and Working Groups

Subcommittee

Subcommitee Steering Group

Kerstin Kleese van Dam, Chair, *Brookhaven National Laboratory*
Adam Schlosser, Co-Chair, *Massachusetts Institute of Technology*
Jeremy Schmutz, Co-Chair, *HudsonAlpha Institute for Biotechnology*
Ben Bond-Lamberty, *Pacific Northwest National Laboratory*
Kjiersten Fagnan, *Lawrence Berkeley National Laboratory*
Ann M. Fridlind, *National Aeronautics and Space Administration*
Susan Gregurick, *National Institutes of Health*
Dev Niyogi, *University of Texas, Austin*
Daniel Segrè, *Boston University*
Pamela Weisenhorn, *Argonne National Laboratory*

Subcommitee Members

Steve Allison*, University of California–Irvine*
Ben Blaiszik*, University of Chicago*
Casey Burleyson*, Pacific Northwest National Laboratory*
Sen Chiao*, Howard University*
Emiley Eloe-Fadrosh, *Lawrence Berkeley National Laboratory*
Gannet Hallar, *University of Utah*
Chris Henry*, Argonne National Laboratory*
Forrest Hoffman, *Oak Ridge National Laboratory*
Shantenu Jha, *Rutgers University*
Ravi Madduri, *Argonne National Laboratory*
Jennie Rice*, Pacific Northwest National Laboratory*
Ratna Saripalli*, Pacific Northwest National Laboratory*
Shin-Han Shiu, *Michigan State University*
Carlos Soto, *Brookhaven National Laboratory*
Michela Taufer, *University of Tennessee*
Luke van Roekel, *Los Alamos National Laboratory*
Charu Varadharajan, *Lawrence Berkeley National Laboratory*
Lou Woodley, *Center for Scientific Collaboration and Community Engagement*

Working Groups

Environmental Science

Ann M. Fridlind, Co-Lead, *National Aeronautics and Space Administration*
Ben Bond-Lamberty, Co-Lead, *Pacific Northwest National Laboratory*
Alison Appling, *U.S. Geological Survey*
Bill Collins, *Lawrence Berkeley National Laboratory*
Gannet Hallar, *University of Utah*
Jessica Haskins, *University of Utah*
Julie McClean, *Scripps Institution of Oceanography*
Daniel McCoy, *University of Wyoming*
Jennie Rice, *Pacific Northwest National Laboratory*
Roy Rich, *Smithsonian Environmental Research Center*
Scott Rupp, *University of Alaska Fairbanks*
Alexey Shiklomanov, *National Aeronautics and Space Administration*
Cove Sturtevant, *National Ecological Observatory Network*
Ning Sun, *Pacific Northwest National Laboratory*
Gunilla Svensson, *Stockholm University*
Luke van Roekel, *Los Alamos National Laboratory*
Charu Varadharajan, *Lawrence Berkeley National Laboratory*

Biological Science

Daniel Segrè, Co-Lead, *Boston University*
Pamela Weisenhorn, Co-Lead, *Argonne National Laboratory*
Frank Alexander, *Brookhaven National Laboratory*
Steve Allison, *University of California, Irvine*
Gorgy Babnigg, *Argonne National Laboratory*
Bruno Basso, *Michigan State University*
Emiley Eloe-Fadrosh, *Lawrence Berkeley National Laboratory*
Pubudu Handakumbura, *Pacific Northwest National Laboratory*
Chris Henry, *Argonne National Laboratory*
Adina Howe, *Iowa State University*
Jeffrey Kimbrel, *Lawrence Livermore National Laboratory*
Andrew Leakey, *University of Illinois Urbana-Champaign*
Jeremy Schmutz, *HudsonAlpha Institute for Biotechnology*
Nadia Shakoor, *Donald Danforth Plant Science Center*
Shin-Han Shiu, *Michigan State University*
Eva Sinha, *Pacific Northwest National Laboratory*

Diversity, Equity, Inclusion, and Accessibility

Pamela Weisenhorn, Co-Lead, *Argonne National Laboratory*
Dev Niyogi, Co-Lead, *University of Texas, Austin*

Hamed Alemohammad, *Clark University*
Jean Andino, *Arizona State University*
Sen Chiao, *Howard University*
Lesley-Ann Dupigny-Giroux, *University of Vermont*
Ward Fisher, *University Corporation for Atmospheric Research*
Joseph Graves, *North Carolina Agricultural and Technical State University*
Ruby Leung, *Pacific Northwest National Laboratory*
Nirav Merchant, *University of Arizona*
Rahul Ramachandran, *National Aeronautics and Space Administration*
Hanadi Rifai, *University of Houston*
Jamese Sims, *Mississippi State University*
Roselyne Tchoua, *DePaul University*
Tanya Vance, *University Corporation for Atmospheric Research*
Lou Woodley, *Center for Scientific Collaboration and Community Engagement*
Elisha Wood-Charlson, *Lawrence Berkeley National Laboratory*

BER Data Services

Casey Burleyson, Co-Lead, *Pacific Northwest National Laboratory*
Kjiersten Fagnan, Co-Lead, *Lawrence Berkeley National Laboratory*
Adam Arkin, *Lawrence Berkeley National Laboratory*
Kristofer Bouchard, *Lawrence Berkeley National Laboratory*
Shreyas Cholia, *Lawrence Berkeley National Laboratory*
Danielle Christianson, *Lawrence Berkeley National Laboratory*
Robert F. Fischetti, *Argonne National Laboratory*
Héctor García Martin, *Lawrence Berkeley National Laboratory*
Forrest Hoffman, *Oak Ridge National Laboratory*
Daniel Jacobson, *Oak Ridge National Laboratory*
Lee Ann McCue, *Pacific Northwest National Laboratory, Environmental Molecular Sciences Laboratory*
Giri Prakash, *Oak Ridge National Laboratory*
Ratna Saripalli, *Pacific Northwest National Laboratory*
Laurie Stephey, *Lawrence Berkeley National Laboratory*
Leslie Stoecker, *University of Illinois Urbana-Champaign*
Jason Zurawski, *Lawrence Berkeley National Laboratory*

Unified Data Infrastructure and Artificial Intelligence

Susan Gregurick, Co-Lead, *National Institutes of Health*
Kerstin Kleese van Dam, Co-Lead, *Brookhaven National Laboratory*
Rachana Ananthakrishnan, *University of Chicago*
Laura Biven, *National Institutes of Health*
Ben Blaiszik, *University of Chicago*
Judy Hill, *Lawrence Livermore National Laboratory*
Shantenu Jha, *Rutgers University*
Ravi Madduri, *Argonne National Laboratory*
Ratna Saripalli, *Pacific Northwest National Laboratory*
Adam Schlosser, *Massachusetts Institute of Technology*
Carlos Soto, *Brookhaven National Laboratory*
Arjun Shankar, *Oak Ridge National Laboratory*
Michela Taufer, *University of Tennessee*

Contents

Executive Summary

The Biological and Environmental Research (BER) program within the U.S. Department of Energy (DOE) Office of Science supports large-scale data generation efforts across its two divisions: Biological Systems Science and Earth and Environmental Systems Sciences. These efforts include user facilities in atmospheric radiation measurements, genomics, metabolomics, proteomics, compute, and imaging. In addition, BER supports the development of plant-based fuels; research in biosystems design, environmental microbiomes, and atmospheric systems; energy flux monitoring; climate-based ecosystem experiments; pathogen biopreparedness; and modeling of climate, urban interfaces, and interactions between people and energy resources. For data access, BER supports community data services at its user facilities, along with specialized data initiatives for Earth and environmental science, climate models, genomic and microbial analysis, and multisector dynamics modeling.

Charge and Approach

In October 2022, the BER advisory committee (BERAC) received a charge letter (see p. ii) from the DOE Office of Science director requesting a review of existing capabilities in data management and infrastructure relevant to BER science. The charge also requested a recommended strategy for next-generation data management and analysis within a unified framework. Further goals included identifying new science opportunities that could be enabled by increased integration of BER's facilities while considering advances in artificial intelligence and machine learning (AI/ML). The charge asked BERAC to examine synergistic investments within DOE and at other agencies and the impact of a more unified data infrastructure on the scientific workforce. To address these goals, the appointed subcommittee established five working groups focusing on (1) environmental science; (2) biological science; (3) diversity, equity, inclusion, and accessibility; (4) BER data services; and (5) unified data infrastructure and artificial intelligence. The subcommittee organized a two-day virtual community workshop that included discussions on new unified data infrastructure-enabled science opportunities, barriers to broader inclusion of minorities, support for early career scientists, and potential unified data infrastructure solutions for BER.

Science Opportunities

Earth's environment is rapidly changing on a global scale. The tremendous range of spatial and temporal scales within climate, natural, and human systems complicates efforts within the environmental science field to improve predictive tools and better evaluate mitigation and adaptation strategies. Building tomorrow's integrative tools in environmental science presents an unprecedented challenge to the interoperability of data records required to advance and constrain model development and performance. Progress is limited in Earth systems science by data limitations, including the need to co-locate and use increasingly large datasets and complex models, access private or protected data on human resource use, and accelerate the rate at which new field data can be quality-controlled and distributed. The next generation of predictive models also will require improved data infrastructure in cross-agency storage and computing, unified metadata conventions, and centralized metadata searching.

Connecting biology to Earth system models is necessary to predict future climate impacts on ecosystems and human societies. Ecosystem processes are inherently multiscale. Linking genes in individual species to collections of organisms exchanging nutrients in feedback loops requires multiscale network models that describe the underlying biological processes. Such models may examine, for example, how microbial communities form, how to efficiently grow plants for bioproducts under a varying climate, and how to link plant and microbial processes to impacts in Earth system models. Limitations to building these models include the need to use multidisciplinary datasets that are not yet widely or publicly available. Data discovery, data integration, and knowledge for manipulating these datasets are additional challenges arising from the lack of standardized, accessible workflows as well as systemic issues such as the lack of incentives (funding and recognition) for data sharing and interoperability.

Several science opportunities emerged during subcommittee discussions and workshop breakouts. These crosscutting opportunities apply across much of BER's biological and environmental sciences and represent areas in which the program can significantly impact efforts to understand and manipulate natural and managed resource systems to meet DOE goals. The first opportunity is to leverage decades of collected environmental and biological data to predict biological systems under realistic field conditions at multiple scales. A second opportunity is to develop and test multiscale models that incorporate environment-dependent biological system variables in high resolution. A final opportunity is to make BER data accessible, inclusive, and usable by the broader community. Addressing these challenges will require DOE to encourage shareable, coordinated data collection with standards for field data; develop curated, standardized, and open datasets that can enable multiscale modeling; and make field, environmental, variation, and climate data accessible to support diversity, equity, and inclusion goals.

Current Accessibility of Unified Data Infrastructures

Truly accessible data requires efforts beyond describing a dataset in a publication and making it downloadable. Different expertise is needed to integrate data at different scales, and a challenge for BER is to provide data in a way that enables a non-domain expert to subset these data and perform analyses.

Training for data generators is needed to improve metadata and documentation, and a data governance plan is required to ensure implementation of these improvements. This can be complicated by different communities' data needs, so computational crosswalks and conversion tools are necessary to connect disparate data types. Community-driven, domain-specific standards, combined with additional incentives, may be needed to encourage small data producers to contribute to a unified data infrastructure. In addition, direct support, training, and tools are needed to extend unified data infrastructure use beyond major facilities to universities and minority-serving institutions (MSIs). This extension could be achieved by increasing tool and data awareness, conducting targeted outreach to diverse stakeholders, and supporting mentoring that expands user reach and diversity.

Finally, the utility of a unified data infrastructure system could be greatly increased by having a single point of entry that supports access across multiple data repositories co-located with computational tools to handle the data volumes.

Workforce Development

Unified data infrastructure can play a critical role in expanding BER's workforce by increasing early career scientists' access to and use of data for scientific advances. One opportunity is to create for diverse communities interfaces that support place-based inquiry and to develop training on standardized sample collection and processing, which could engage broader participation in a unified data infrastructure. These efforts would enable real-time data sharing and collaboration and help mitigate the impacts of the constantly changing state of the art for analyses. Additional workshops are needed to address how best to incentivize and accelerate participation at underfunded schools, including infrastructure improvements and direct funding opportunities. These efforts and their development may require extensive and inclusive trust building as well as integrating into the decision-making processes the faculty from underfunded schools such as community colleges, historically black colleges and universities (HBCUs), and other MSIs.

Data Infrastructure and AI

Biological and environmental sciences increasingly require the integration of multidisciplinary and multiscale datasets, and specific challenges vary by community. In environmental science, data is usually available and accessible, but, in many cases, numerous exceedingly large datasets are required for integrated research. Across the biological sciences, datasets, while more manageable, are often sparse, incomplete, and inaccessible to the broader research community.

Integrating data from multiple sources into one coherent body of information for further analysis remains a predominantly manual task. Repeated downloads, transfers, and processing of numerous large datasets are untenable for most users and disproportionately affect those at institutions with fewer resources. AI use is currently limited, and numerous technical barriers to entry may make existing tools unsuitable for noncomputing experts. Furthermore, many of these tools are platform-specific, lack community support, or are partially or fully closed-source in some cases. These limitations may preclude the use of these tools in a broad, public, and multi-institutional integrated infrastructure.

Nevertheless, several promising granular AI solutions could be broadly adopted, customized, implemented, and learned. To achieve at-scale advances in AI-supported capabilities, data policies that support standardization and integration are essential for enforcing inclusive, community-wide governance structures. The recent White House Executive Order on AI adds additional requirements to ensure that deployed AI solutions are safe, secure, and trustworthy (U.S. White House 2023). Educating and distributing best practices and tests to the BER community could best be achieved through a unified data infrastructure that democratizes access to these tools. Additional development requirements include scalable search engines and a common mechanism for attributions. Also needed, but not currently supported by existing BER data facilities, is a collaborative data fabric that ties facilities to a standard user experience and interface and includes metadata, ontologies, workflows, clean and ready-to-use datasets, tools, and frameworks for storing data and metadata from small and large producers. All these advances must be paralleled with strong and integrated training, support, and outreach programs to minimize entry barriers and change current working practices.

Summary and Recommendations

Although BER has a sophisticated set of data infrastructure capabilities, the subcommittee identified gaps in data support and accessibility. For example, there is limited support for cross-community integration of data types, connections with other agencies' data services, and tool sharing. Such capabilities are critical since BER research is increasingly complex and requires the integration and study of processes across scales and modalities. Current BER data infrastructure is not ready to support such efforts, however. In addition, increased effort could help underserved communities, minorities, and early career scientists more easily access BER capabilities and participate in BER research.

Development of a BER unified data infrastructure could involve learning from existing worldwide efforts. It also would require technical innovation and the integration of researchers from different communities into an infrastructure enabling them to communicate and interact with ease. A complete solution is not expected to be achievable in 5 years, but the subcommittee identified multiple steps that can serve as critical stepping stones with tangible, community-oriented scientific benefits. Specific subcommittee strategic recommendations follow.

- Include researchers and developers when developing the infrastructure and explicitly target outreach to diverse stakeholders, including mentoring to promote awareness.
- Identify a select number of high-impact science goals that require a unified data infrastructure and ultimately affect a culture change across the BER research space.
- Leverage existing BER facilities and data services to build an initial tightly integrated unified data infrastructure. Augment this infrastructure with a dedicated data facility (can be federated) that combines large-scale data and computing to alleviate the need for BER scientists to download data for integration and analysis.

- Establish a BER "marketplace" where BER scientists can discover and use data, tools, services, and resources across BER programs; form new collaborations; and regularly evolve the infrastructure as the community changes.
- Co-develop plans to leverage data infrastructures, such as the (1) DOE Advanced Scientific Computing Research program's Integrated Research Infrastructure High-Performance Data Facility; (2) the National Science Foundation's National Scientific Data Fabric; and (3) European Open Science Cloud efforts, including the European Destination Earth.
- Support select integration and interaction with other agencies' data frameworks that are important to BER science and that can accelerate the unified data effort.
- Support the integration—at scale—of new technologies such as AI, quantum, and digital twins, along with transformative open data policies that facilitate needed expansion and development of existing granular solutions. For all these capabilities, offer integrated training, education, and outreach programs.

1 Introduction

1.1 BER Mission and Research Portfolio

The Biological and Environmental Research (BER) program within the U.S. Department of Energy's (DOE) Office of Science stewards transformative research and scientific user facilities to achieve a predictive understanding of complex biological, Earth, and environmental systems at scales ranging from molecules to the whole planet. BER supports fundamental research into the relationships between energy and environment to create the foundation for a sustainable and reliable energy future. The program's research is organized into two divisions: the Biological Systems Science Division (BSSD) and the Earth and Environmental Systems Sciences Division (EESSD).

BSSD supports fundamental science to understand, predict, manipulate, and design biological systems that underpin innovations for bioenergy and bioproduct production and to enhance the understanding of natural, DOE-relevant environmental processes (U.S. DOE 2021). Within its systems biology portfolio, BSSD supports genomic science, proteomics, metabolomics, structural biology, computational modeling, and bioimaging research and the application of these approaches to plants, microbes, and communities.

EESSD supports research to characterize and understand feedbacks between Earth and energy systems, including studies on atmospheric physics and chemistry, ecosystem ecology, and biogeochemistry. The division also supports efforts to develop, validate, and analyze Earth system models that integrate information on the biosphere, atmosphere, terrestrial land masses, oceans, sea ice, land ice, the subsurface, and human components to advance scientific understanding and improve Earth system predictability.

1.2 BER-Relevant User Facilities and Capabilities

The BER research community leverages a network of large-scale DOE Office of Science user facilities, some funded by BER and some by other programs, that create significant volumes of research data:

- **BER's Atmospheric Radiation Measurement (ARM) User Facility, *www.arm.gov.*** ARM offers highly instrumented ground stations at various locations around the globe, mobile measurement resources, and aerial vehicles to continuously measure cloud and aerosol properties and their impacts on Earth's energy balance. ARM measurements have set the standard for long-term climate research observations and provide an unparalleled resource for examining atmospheric processes and evaluating Earth system model performance.
- **BER's Joint Genome Institute (JGI), *jgi.doe.gov*.** The JGI user facility sequences more than 450 trillion DNA bases per year and provides state-of-the-science capabilities for genome sequencing, synthesis, metabolomics, and analysis. With nearly 1,600 users worldwide on active projects, JGI is the preeminent resource for sequencing plants, fungi, algae, microbes, and microbial communities foundational to energy and environmental research.
- **BER's Environmental Molecular Sciences Laboratory (EMSL), *www.emsl.pnnl.gov*.** EMSL provides users with a problem-solving environment by integrating premier instrumentation with high-performance computing and optimized codes. This integration of capabilities

enables research teams or individual investigators to unravel the fundamental physical, chemical, and biological mechanisms and processes that underpin larger-scale biological, environmental, and energy challenges.

- **Advanced Scientific Computing Research (ASCR) Program's Leadership Computing Facilities, *science.osti.gov/User-Facilities/User-Facilities-at-a-Glance/ASCR/*.** These facilities provide peta- and now exascale computing capabilities enabling detailed computational modeling efforts. For BER researchers in particular, these efforts include climate science modeling such as the Energy Exascale Earth System Model (E3SM) and smaller modeling efforts in biological, Earth, and environmental science. More recent artificial intelligence (AI) efforts are similarly supported by ASCR facilities.
- **Basic Energy Sciences (BES) Program's and BER's Imaging Capabilities, *berstructuralbioportal.org*.** Housed at BES synchrotron and neutron facilities, BER-supported technologies and measurements in structural biology and imaging can resolve key metabolic processes over time within or among cells. These capabilities, including cryo-electron microscopy resources, are supported by BER's Biomolecular Characterization and Imaging Science program and produce information that serves as a crucial bridge toward linking molecular-scale information to whole-cell, systems-level understanding.

1.3 Large-Scale Data-Generating Projects

In addition to user facility–based capabilities, BER funds a range of large-scale projects and programs that create or use significant or diverse volumes of data. These include, but are not limited to, the following:

- **The Bioenergy Research Centers (BRCs), *genomicscience.energy.gov/bioenergy-research-centers/*.** The mission of the BRC program is to break down the barriers to actualizing a domestic bioenergy industry. The four centers—each led by a DOE national laboratory or top university—take distinctive approaches toward the common goal of accelerating the pathway to improving and scaling up advanced biofuel and bioproduct production processes.
- **AmeriFlux, *ameriflux.lbl.gov*.** Supported by BER and the National Science Foundation, AmeriFlux is a network of principal investigator–managed sites measuring ecosystem carbon dioxide, water, and energy fluxes in North, Central, and South America. It was established to connect research on field sites representing major climate and ecological biomes, including tundra; grasslands; savanna; crops; and conifer, deciduous, and tropical forests.
- **Next-Generation Ecosystem Experiments (NGEEs), *ess.science.energy.gov/critical-ecosystems/*.** The multiphased NGEE projects aim to improve predictive understanding of specific climate zones such as the Arctic and tropics. This objective is achieved through experiments, observations, and synthesis of existing datasets that strategically inform model process representations and parameterizations and that enhance the knowledgebase required for model initialization, calibration, and evaluation.
- **Biopreparedness Research Virtual Environment (BRaVE), *science.osti.gov/Initiatives/Biopreparedness/*.** The BraVE program aims to address a range of potential biological events and transform the nation's ability to prepare for and respond to future biological threats. Current multidisciplinary projects, supported by ASCR, BES, and BER, are also seeking to provide broader insights into fundamental processes in biological systems and develop new characterization and computational approaches relevant to DOE research in renewable energy, climate change, biomanufacturing, and the broader bioeconomy.
- **Global Change Analysis Model (GCAM), *gcims.pnnl.gov/modeling/gcam-global-change-analysis-model/*.** GCAM is a market equilibrium community model with a global scope. Other

socioeconomic models include those that follow a computable general equilibrium approach. All these models are designed to study how changes in population, income, or technology cost would be expected to alter crop production, energy demand, and water use throughout an interconnected global environment.

- **E3SM, *e3sm.org*.** Using exascale computing, this project conducts high-resolution Earth system modeling of natural, managed, and man-made systems to answer pressing DOE mission challenges. In particular, researchers are using E3SM to study long-term trends that will have major impacts on the energy sector. Among these are regional trends in air and water temperatures, water availability, storms and heavy precipitation, coastal flooding, and sea-level rise. The ability to simulate and predict significant, long-term global changes is important to energy-sector and policy-relevant planning.
- **Urban Integrated Field Laboratories (UIFLs), *ess.science.energy.gov/urban-ifls/about/*.** These multidisciplinary projects are advancing the science that underpins the predictability of urban systems and their two-way interactions with the climate system. The four UIFLs also aim to provide the knowledge and information necessary to inform equitable climate and energy solutions that can strengthen community-scale resilience across urban landscapes.

1.4 BER Community Data Services

To preserve and make available its wealth of research data, BER has created a range of community-based data services that enable the co-location and, where appropriate, the integration of BER data across projects and within targeted communities. These include:

- **ARM, JGI, and EMSL Data Services**. To make data collected at their respective facilities more broadly available, these facilities offer data services to their communities.
  - *www.arm.gov/connect-with-arm/organization/data-services*
  - *jgi.doe.gov/data-and-tools/*
  - *www.emsl.pnnl.gov/data-management-policy/*
- **Environmental System Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE), *ess-dive.lbl.gov*.** ESS-DIVE is a data repository for Earth and environmental sciences. It collects, stores, manages, and shares environmental systems data created through BER-sponsored research, including the NGEEs and AmeriFlux.
- **Earth System Grid Federation (ESGF), *esgf.llnl.gov*.** Led by BER, ESGF is an international collaboration that develops, deploys, and maintains software infrastructure for managing, disseminating, and analyzing climate model output and observational data. E3SM uses ESGF, which is also supported by other U.S. and international sponsors.
- **DOE Systems Biology Knowledgebase (KBase), *kbase.us*/.** KBase is a knowledge creation and discovery environment designed for biologists and bioinformaticians. It allows users to perform large-scale analyses and combine multiple lines of evidence to model plant and microbial physiology and community dynamics.
- **National Microbiome Data Collaborative (NMDC), *microbiomedata.org*.** NMDC is enabling inclusive and interdisciplinary environmental microbiome science by connecting data, people, and ideas. The project's scientific mission is to provide comprehensive discovery of and access to multiomics microbiome data.
- **MultiSector Dynamics–Living Intuitive Value-adding Environment (MSD-LIVE), *msdlive.org*.** MSD-LIVE is a flexible and scalable data and code management system combined with a distributed computational platform that will enable MSD researchers to document and archive their data; run their models and analysis

tools; and share their data, software, and multimodel workflows.

1.5 BERAC Charge on Unified Data Infrastructure

As described above, BER has a wide-ranging and influential portfolio of data resources that in themselves form unified data infrastructure islands that serve specific communities (see Fig. 1.1, p. 5). However, not all science projects and facilities have an associated data service. For example, the bioimaging capabilities and BRCs are not covered by any of the existing data services. Furthermore, none are integrated with each other, resulting in customized or duplicated metadata, data management, access protocols, and analysis services.

To address these challenges, the BER advisory committee (BERAC) received a charge from the DOE Office of Science director in October 2022 requesting a review of existing capabilities in data management and infrastructure relevant to BER science.

The charge letter asked BERAC to (1) review the existing and anticipated capabilities in data management and supporting infrastructures that are relevant to the breadth of BER science and (2) recommend a strategy for the next generation of data management and analysis within a unified framework.

In addition, BERAC was asked to:

- Identify new science opportunities that could be possible within and across BER programs if a unified data framework were to be developed.
- Assess recommendations from recent AI and machine learning reports that potentially could be incorporated into a future BER data framework (e.g., with a component that includes training data).
- Consider data management strategies and investments in other agencies that could be leveraged in developing a BER unifying framework.
- Provide a list and brief explanation of the components and specifications that would be needed to develop in the next 5 years a unified framework for BER science.
- Examine how developing a unified data framework would benefit the scientific research workforce, with particular attention to increased opportunities for enhancing career progression and which types of culture changes could help facilitate those benefits.

1.6 Approach to Addressing the Charge

To address these objectives, BERAC took the following actions.

- Established a subcommittee with five working groups focused on (1) environmental science; (2) biological science; (3) diversity, equity, inclusion, and accessibility; (4) BER data services; and (5) unified data infrastructure and artificial intelligence.
- Expanded the workforce topic to include underserved communities and minorities in inclusion and accessibility related to existing and future BER data resources.
- Expanded the workforce topic to incorporate support for early career researchers in BER.
- Reviewed current BER data infrastructures and activities that could lead to closer integration of existing resources.
- Reviewed worldwide state-of-the-art efforts in creating and operating unified data infrastructures.
- Released a Request for Information (RFI) to the BER community to comment and provide input to the charge questions above.
- Organized a two-part community workshop where day one focused on (1) identifying new science opportunities that could be enabled by a unified data infrastructure and (2) identifying current barriers to realizing these science opportunities. On day two, teams discussed barriers to

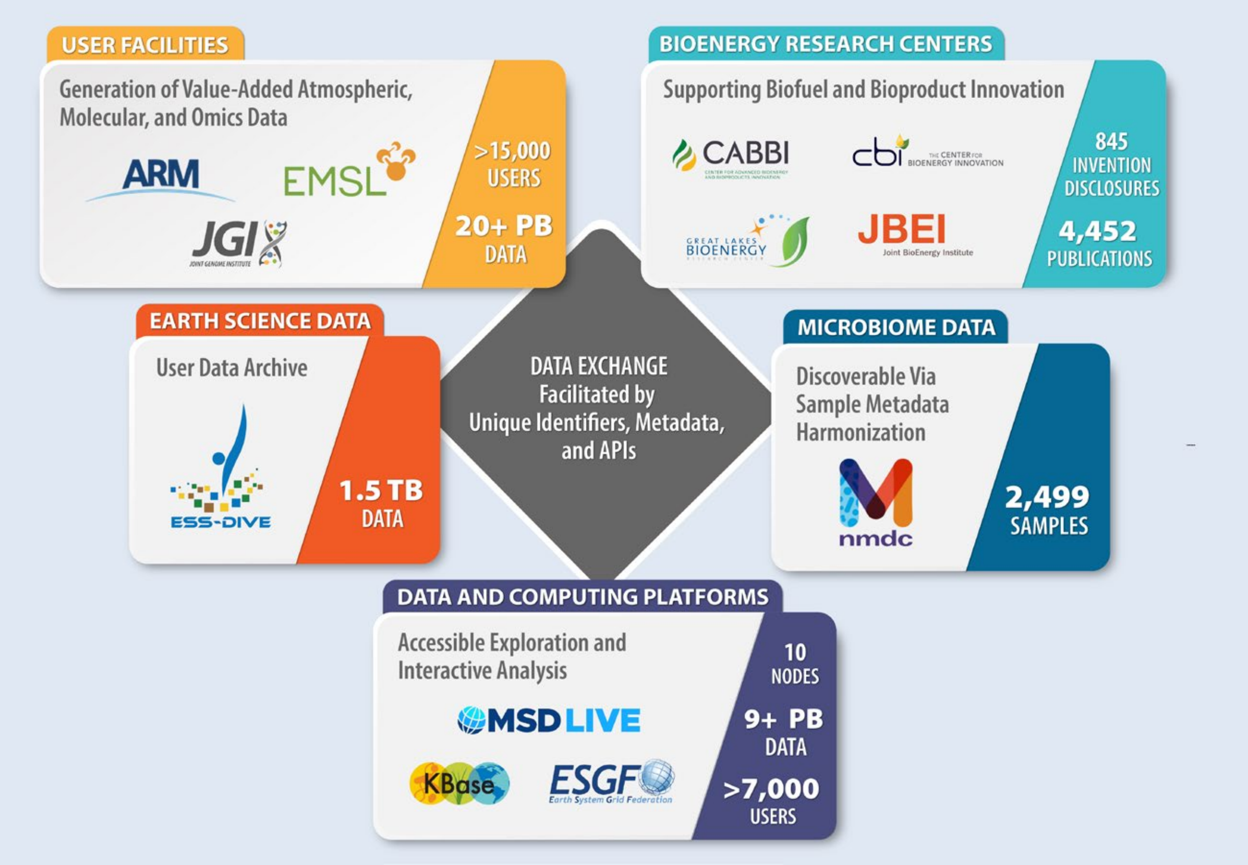

USER FACILITIES
Generation of Value-Added Atmospheric, Molecular, and Omics Data
ARM
EMSL
JGI
>15,000 USERS
20+ PB DATA
BIOENERGY RESEARCH CENTERS
Supporting Biofuel and Bioproduct Innovation
CABBI
cbi THE CENTER FOR BIOENERGY INNOVATION
GREAT LAKES BIOENERGY
JBEI Joint BioEnergy Institute
845 INVENTION DISCLOSURES
4,452 PUBLICATIONS
EARTH SCIENCE DATA
User Data Archive
ESS-DIVE
1.5 TB DATA
DATA EXCHANGE Facilitated by Unique Identifiers, Metadata, and APIs
MICROBIOME DATA
Discoverable Via Sample Metadata Harmonization
nmdc
2,499 SAMPLES
DATA AND COMPUTING PLATFORMS
Accessible Exploration and Interactive Analysis
MSD LIVE
KBase
ESGF Earth System Grid Federation
10 NODES
9+ PB DATA
>7,000 USERS

Fig. 1.1. BER Infrastructure for Synergistic Data Generation, Exploration, and Integrative Analysis. 2023 statistics for BER user facilities and resources. [Courtesy Lawrence Berkeley National Laboratory]

broader inclusion of minorities and strategies to better support early career scientists. Participants also discussed possible unified data infrastructure solutions to address the barriers identified on day one and two.

To develop this report, subcommittee members reviewed the rich material collected during the workshop and through the RFI and synthesized their findings pursuant to the charge. Ch. 2, p. 7, describes current data infrastructure barriers across BER's research portfolio and seeks to answer the question: what scientific opportunities would be enabled if such barriers could be lowered through a better unified data infrastructure? Ch. 3, p. 23, reviews the current state-of-the-art in inclusion and accessibility for unified data infrastructures and outlines barriers and approaches to overcome them. Ch. 4, p. 33, reviews the ability of existing state-of-the-art efforts to address these barriers and identifies specific actions for the future. Finally, Ch. 5, p. 43, describes the subcommittee's high-level recommendations in detail.

2 Science Opportunities

The overriding objective for new BER investment in data unification is advancement of BER-supported scientific endeavors. Achieving BER and national priorities clearly will require simultaneous innovations in data services that span every aspect of accessibility, along with petascale and exascale scientific data generation, and effective use of such data.

Two examples of pioneering BER projects in data unification that have grown to include extensive contributions internationally are the Earth System Grid Federation's (ESGF) Climate Model Intercomparison Project (CMIP) data repository and the AmeriFlux network. ESGF hosts modeling data contributions worldwide, and AmeriFlux supports observational data contributions from ground sites managed by individual users throughout North, Central, and South America.

Both projects represent pivotal and sustained DOE investments, and their use by the U.S. and international scientific communities continues to be widespread as self-supported users contribute to the projects free of charge. Contributions, in turn, adhere to a particular set of data and site standards that support sustained data integration and accessibility. For example, the BER-supported Program for Climate Model Diagnosis and Intercomparison at Lawrence Livermore National Laboratory contributed to establishing the international Climate and Forecast metadata conventions followed by ESGF. This activity demonstrates BER success in foundational contributions to the most widely used community metadata convention in climate science.

As the programs grew, innovations emerged within both ESGF and AmeriFlux. One example is AmeriFlux's establishment of a loan program for calibration instruments. These innovations were not without cost; as programs and data repositories grew, so did expenses, which have been offset by external sponsors in the case of ESGF. However, in both cases, the stewardship of essentially crowdsourced archives of modeling data (ESGF) and observational data (AmeriFlux) created unique conditions for unleashing scientific potential through impactful and sustained community buy-in and contributions that continue to produce groundbreaking scientific results today.

As DOE considers new data unification investments, appropriate questions include: (1) what are the current main barriers to progress experienced by BER data providers and users, and (2) what new scientific avenues would be opened by lowering those barriers? In the following sections, writing team members, themselves scientists working in BER research areas, summarize input on these questions obtained from the Request for Information (RFI) and workshop sessions.

This chapter describes specific opportunities in environmental science (Section 2.1, this page), biological science (Section 2.2, p. 14) and crosscutting science (Section 2.3, p. 19) that could be enabled by a unified data infrastructure and highlights barriers that prevent sufficient progress today.

2.1 Environmental Science Opportunities

Earth's environment is changing rapidly at the global scale, with diverse impacts increasingly observed across ocean, land, and polar regions. Efforts to improve predictive tools and better evaluate mitigation strategies are challenging the entire environmental science field. In particular, there is a critical need to integrate understanding across climate system components (e.g., oceans, land, cryosphere, and biosphere) and life-sustaining human systems (e.g., energy production, urban and transport infrastructure, and agriculture and food production). Since climate and human systems span a tremendous range of spatial and temporal scales, integrated analysis tools demand unprecedented access to measurement and model data in every sense of the term.

This section deliberately avoids a narrow interpretation of "environmental science," acknowledging that future challenges will be integrative in nature. Instead, an open-ended definition is used. This definition includes not only the traditional areas of climate, hydrology, and ecosystem sciences where human activities are often considered a boundary condition but also multisector dynamics and urban environments. Also included in the definition is a future range of application-specific digital twins, such as those planned by the European Destination Earth project (see Fig. 2.1, this page).

2.1.1 What Research Could a Unified Data Infrastructure Enable?

New measurement capabilities are often said to drive new science. However, building tomorrow's integrative tools in the environmental sciences presents an unprecedented challenge to both national and international interoperability of data records required to advance and constrain model physics and performance. Based on RFI and workshop responses, BER scientists generally are taking on this challenge in a piecemeal fashion because their progress depends on it. However, their current tools and workflows frequently face limitations, which, if removed, would accelerate scientific advances. Workshop participants identified two specific scientific advances that could be enabled if a unified data infrastructure existed: multisector dynamics and Earth and environmental system modeling.

Multisector Dynamics

In the multisector dynamics field, researchers require data from NASA, the National Oceanic and Atmospheric Administration (NOAA), U.S. Geological Survey (USGS), Environmental Protection Agency

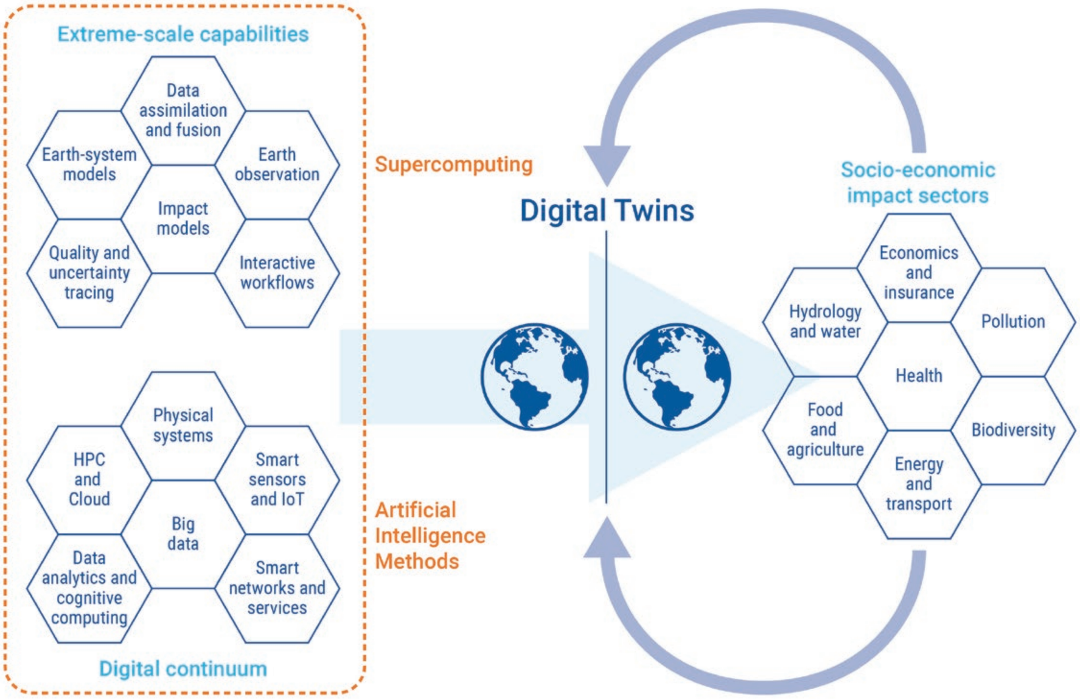


Fig. 2.1. The European Destination Earth Project: An Example of Integrated Forecasting and Policy Tools. The project will rely on a core service platform and distributed data lake, whose management will be led by the European operational satellite agency (EUMETSAT). This report's writing team is unaware of such plans at the U.S. federal level. [Reused under a Creative Commons license (CC BY 4.0) from "Destination Earth," © European Union, digital-strategy.ec.europa.eu/en/library/destination-earth. Colors modified from original for this report.]

(EPA), and the U.S. Forest Service (USFS). However, these data currently are not linked with sociodemographic, health, infrastructure, economic, trade, or population data in ways that would facilitate analysis across natural, managed, and built environments. Integrating high-resolution urban-scale field measurements would provide a foundation for advancing research on urban vulnerability and resilience to climate and nonclimate stressors.

To develop Earth system models, researchers need to use ground-based measurements from BER's Atmospheric Radiation Measurement (ARM) facility together with NASA and NOAA satellite measurements to reduce uncertainty in climate model physics in a methodologically sound and robust fashion. Surmounting this barrier would significantly speed up physics development in areas that are currently limiting the confidence in predictions of transient and equilibrium climate sensitivity (see Box 2.1, this page). In addition, linking AmeriFlux and ESS-DIVE data to NASA, NOAA, and USGS data would enable generalization of BER data, helping to constrain carbon cycle dynamics regionally.

Oceans store and move vast quantities of anthropogenically generated heat via processes such as the global thermohaline circulation. Critical uncertainties in future ocean behaviors exist due to both difficulties in integrating sparse observations and challenges associated with enormous data volumes emerging from high-resolution modeling. Across surface hydrology, ocean science, and polar science (e.g., permafrost modeling and river ice), artificial intelligence and machine learning (AI/ML) approaches offer tremendous promise. However, this potential is limited because projects end up devoting most of their time to data curation, quality control, and formatting. AI could help with such geospatial data fusion by enabling researchers to create their own databases rather than relying on DOE to stand up intensive efforts.

Earth and Environmental System Modeling

In Earth and environmental system modeling, connecting the onset of extreme events to their impacts would enable researchers to better characterize them and project the effects of extremes onto

Box 2.1 Robustly Predicting How the Nonlinear Earth System Will Respond to 21st Century Climate Change

Outcomes: Quantify Earth's transient and equilibrium climate sensitivity under realistic emissions scenarios and reliably predict conditions associated with dangerous extremes and tipping points.

Challenges: (1) Lack of efficient simultaneous access to massive and growing multiagency data sources (e.g., satellite, ground-based, and ocean network measurements). (2) Inconsistent metadata conventions. (3) Lack of computationally intensive data processing capabilities that can operate centrally using shared tools. (4) Global datasets that require bespoke efforts to integrate (e.g., oceanic field campaigns and international data sources).

Unified Data Infrastructure Improvements: (1) Unified, consistent access to data archives maintained by different U.S. agencies at the project level (versus the current practice of projects addressing this on a one-off, local basis). (2) Accessible and sufficient server-side computing capability integrated with up-to-date data archives. (3) Support for shared community tools that integrate diverse data sources.

Future Benefits: Reliable climate projections will foster adequate costing and planning necessary for political and economic stability. Disruptions to society are likely to be catastrophic, requiring major investments that must be based on reliable basic scientific foundations (e.g., knowledge of climate sensitivity measures).

biogeochemical and hydrological cycles. Also, efforts to integrate land management practices, environmental engineering, ecosystem restoration, and conservation within current Earth system models (see Box 2.2, this page) critically depend on recovering older information (pre-1980) to understand the long-term feedbacks of slow-moving soil carbon stocks and climate change driven by carbon dioxide (CO_2).

2.1.2 Gaps: What Limits Progress?

RFI responses varied widely in length but were rather sparse in number, perhaps indicating a somewhat inchoate sense of what could be advocated for or reported as a barrier with an obvious solution. Although the opportunities identified in each response ranged broadly within the BER-supported environmental sciences, the challenges centered around common themes.

Figure 2.2, p. 11, presents a rough accounting of specific data unification challenges described within the RFI responses. Two overarching challenges received almost universal mention: synthesis and data formats. This finding is unsurprising given that the environmental science field as a whole is struggling to synthesize disparate datasets with different formats and that data often span multiple U.S. agencies.

More than half of RFI respondents noted data volume, computational support, and AI/ML demands as specific challenges. More than 25% mentioned metadata; data discovery; data resolution; model data formats; quality controls; subsetting; and access protocols, with many advocating for user-friendly application programming interfaces (APIs). A smaller fraction noted barriers or needs involving data availability, centralized governance, updates, usage rights, digital twins, legacy data, submissions, and privacy.

The challenges summarized in Fig. 2.2, p. 11, can be collectively considered within the context of several general use case limitations that correlate and overlap with all workshop topical areas to some degree. The following brief descriptions of limitations involving data interoperability, access, and availability are drawn from workshop use cases. These types of challenges are commonly and individually faced by climate modeling centers or BER-funded projects attempting to integrate BER data sources. Ideally, data infrastructure solutions would address such challenges at large.

Box 2.2 Incorporate State-of-the-Art Knowledge of Human-Earth System Dynamics in Societal Decision-Making

Outcomes: Guide strategies to save lives and infrastructure in the face of rapidly changing extremes (e.g., river flows, flooding, and heat and cold waves unprecedented in human history).

Challenges: (1) Earth system modeling barriers (see Box 2.1, p. 9) and a lack of data on human activities such as household energy use or agricultural water use. (2) Lack of access to industry data subject to privacy and other protections.

Unified Data Infrastructure Improvements: Earth system modeling improvements (see Box 2.1) as well as new agreements, incentives, and privacy protections that enable access to required data.

Future Benefits: (1) Under increasing societal pressures, allocating taxpayer resources to major mitigation strategies will be provably efficient for the multiple stakeholders affected. (2) Unintended consequences of mitigation strategies will be better predicted.

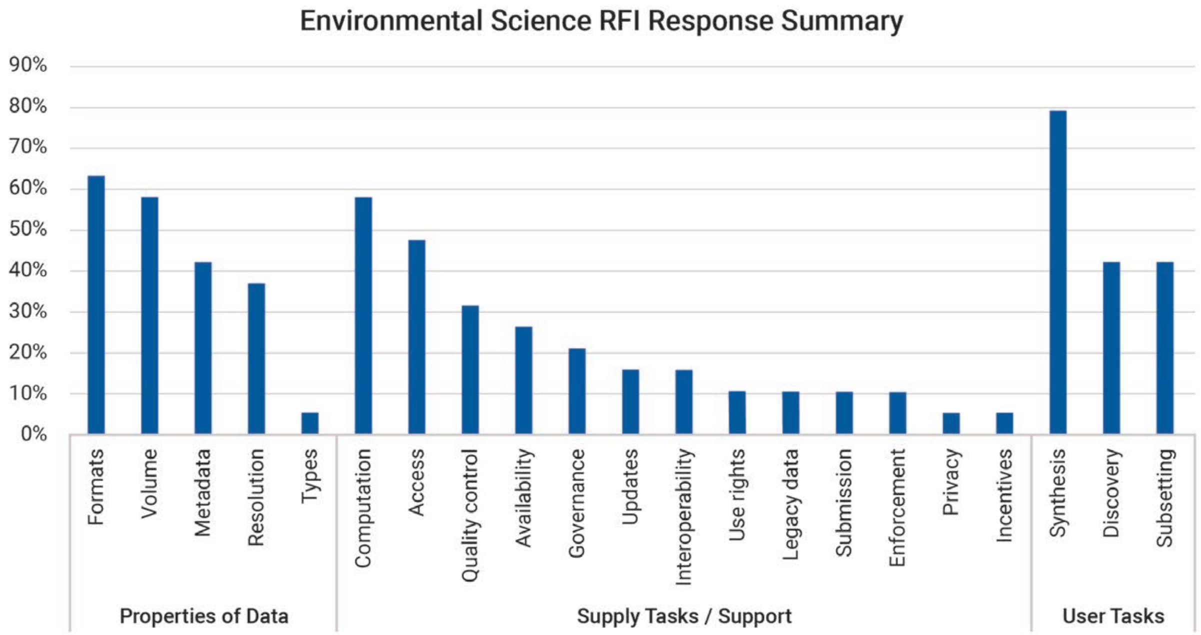


Fig. 2.2 Summary of Data Unification Challenges in Environmental Science. These challenges were identified in written responses to a Request For Information; the percentage of RFI responses that noted each challenge is shown.

Interoperability Limitations—Example: Precipitation at Earth's Surface

Precipitation at Earth's surface is a key quantity for a wide array of environmental science. Major U.S. investments are ongoing in global satellite retrievals and a national ground-based polarimetric radar network. Motivated by uncertainty in climate physics, BER contributes globally distributed surface-based *in situ* and remote-sensing measurements uniquely capable of constraining the lightest rain rates, which are particularly relevant to climate projection. These data are paired with detailed ancillary ground-based measurements.

However, global model developers currently cannot use these data without downloading increasingly large datasets independently, repeatedly, and largely in their entirety. This barrier can be viewed as a synthesis challenge involving data formats, resolution, and huge data volumes subject to rolling updates and revisions. Moreover, differences in metadata and quality controls confound efficient subsetting. In general, most data-hungry environmental applications repeatedly face these challenges, where the need for additional variables requires a long list of distributed resources, which rapidly becomes a barrier to efficient use of the richest datasets.

Access Limitations—Example: Data on Human Activities

Data on human activities, such as household energy use or agricultural water use, is a central need for multisector dynamics, digital twins, and other applications that address the climate-ecosystem-human systems nexus. To some degree, such data exist in industry but are subject to privacy and other protections that limit their use for BER-supported environmental sciences.

Additional examples include global soil dataset products, the Community Emissions Data System (CEDS) containing anthropogenic emissions data on reactive gases and aerosols, the proprietary but widely used International Energy Agency databases, and the Global Land Data Assimilation System. These and many other data at regional to global scales are used by a wide variety of Earth system models, dynamic global vegetation models, and integrated human–Earth system models. For instance, these data provide crucial constraints for integrated assessment and energy systems models, particularly when used in the "target-finding" modes needed to generate economically possible, globally consistent scenarios (e.g., Representative Concentration Pathways and Shared Socioeconomic Pathways) for the international climate change community.

Availability Limitations — Example: Oceanic Surface Layer Data

Oceanic surface layer data is so sparse globally that significant project effort is required to discover and integrate cruise and campaign measurements, which are slow to come online owing to the challenges of normalizing diverse measurement techniques and reporting formats. This limitation poses a crucial challenge to ocean and Earth system modeling, as oceans are the dominant long-term control on atmospheric CO_2 concentrations and the rate of global temperature equilibration under elevated CO_2.

2.1.3 Discussion and Takeaways

RFI respondents and workshop participants reported common data infrastructure challenges in environmental science. DOE scientists are not alone in facing these challenges, as evident in detailed RFI responses submitted by leadership groups from NOAA and USGS, which included presentations, publications, and extensive ancillary information. These RFI responses describe large efforts to overcome these barriers that are being simultaneously mounted across the U.S. government agencies that collectively foster national environmental science, especially those that supply data. Such agency-level responses also indicate a desire to communicate, offer additional information if helpful, and a hope for "interoperability guarantees."

To meet the goal of providing its first ecosystem of digital twins, the European Destination Earth effort can be expected to address many of these major challenges with top-down coordination. Destination Earth's data lake capability will be paired with industrial cloud computing, a common metadata system, and centralized search capabilities. Another expected outcome of this effort is the more rapid identification of specific data availability and existence gaps and how they might be addressed institutionally.

Compared with Europe's centralized Destination Earth effort, U.S. agencies are tasked with coordinating a bottom-up effort that will need to be advanced jointly. How can BER help incubate such a future now? Past data access models might suggest simply bringing two agency data repositories together. For instance, NASA satellite data and ARM site data could be mirrored to a single repository with a joint search capability. This approach would be akin to "build (some of) it, and they will come." However, more far-reaching and general solutions are likely to be those that provide, in coordination with U.S. agency partners, a scalable foundation that attracts many individual efforts en masse.

RFI respondents offered some high-level perspectives on current challenges. One respondent urged lightweight approaches, or "learn as you go" strategies, and emphasized over all else the use of easily programmable APIs accessible to Python, in particular. Another respondent noted that the research community already knows how to build the tools it needs but lacks ways as a community to decide "the why and the shape" of those builds; that respondent urged a focus on governance rather than server space or technical equipment.

The authors of this report broadly agree that while the barriers to unified data solutions are clear, the path forward is not. Effective solutions will need to be coordinated with various efforts external to BER. A danger is premature investment in expensive solutions that may not be generalizable or effective. The subcommittee therefore suggests a project-oriented approach that focuses ideally over a range of currently supported BER projects that face diverse challenges. Pairing a technical expert team with scientists on the ground will ensure positive results from investments in real time. At a project level, results will help major projects move forward. At a meta-information level, pathways forward and barriers can be systematically evaluated together and discussed among agencies to illuminate next steps and directions.

RFI respondents, workshop participants, and co-authors of this report recognize that BER could play a leadership role in the development of a unified data infrastructure. BER has the hardware, the expertise in both technical and community-building, and a uniquely diverse portfolio of data providers and users that is rich with opportunities described in this section and the next. The Destination Earth project provides one model for future efforts at an industrial scale. The pathways to achieving future integration efficiently in the United States likely involve trial and error within the context of agency mandates, priorities, and limitations.

Environmental science project-level trials could usefully be selected to tackle each major class of challenges described in Section 2.1.2, p. 10. If the future is partly envisioned as an interoperable "data archive of data archives," potential targets are myriad. For instance, aircraft campaign data provide a gold standard for aerosol, cloud physics, and trace gas measurements (see Box 2.3, this page). ARM, NASA, and the National Science Foundation (NSF) all house archives of flight campaign data in different formats and with different types of metadata. This archive is likely the most difficult to navigate within BER's data portfolio, according to workshop discussions.

The great potential in integrating U.S. and international aircraft campaign data has been part of community discussions abroad. Such an effort could be supported as a selected pilot project involving, for example, trial efforts to overcome many of the barriers described throughout this report (e.g., metadata conversion, documentation, and interagency coordination). It also could be partly supported as a joint crowd-sourcing effort involving individual principal investigators in the United States and potentially internationally. Other projects could address efficient access to extremely large datasets (e.g., novel grid-averaging approaches used within the global storm-resolving model community) and contextualization of detailed site measurements within satellite datasets (see also Section 2.3, p. 19). The subcommittee suggests the possibility of direct participation from currently

Box 2.3 Fill Gaps in Process-Level Understanding Needed to Improve Multiscale Earth System Model Physics

Outcomes: One example is improved knowledge of the fundamental aerosol and cloud processes that most limit current confidence in historical and future radiative forcing and cloud feedbacks.

Challenges: (1) Difficulty in coordinating data integration and storage across U.S. agency and international participants in major field campaigns. (2) Inconsistent methodological, metadata, and other standards. (3) Inconsistent data deposition by individual principal investigators and laboratories. (4) No coordinated searchability across data available within BER and from U.S. and international partners.

Unified Data Infrastructure Improvements: (1) Standardized formats co-developed by modelers and observationalists. (2) Community-supported expansion of Climate and Forecast conventions when they become inadequate. (3) Interoperability of data from fixed and moving platforms. (4) Community toolkits to create compliant files, check format adherence, and visualize data for common use cases.

Future Benefits: (1) Easier process-based model evaluation and intercomparison will facilitate model validation and improvement. (2) Data that are paradoxically uniquely valuable and underutilized become widely useful by dramatically lowering barriers to access and usability.

funded projects already facing the monumental challenges at hand, thus substantially reducing the danger of expensive errors and paving the way for community-based incubation.

2.2 Biological Science Opportunities

Biological systems play critical roles in (1) transforming the environment; (2) implementing otherwise inaccessible sustainable catalysis for chemicals and materials for the circular bioeconomy; and (3) underpinning the health and biosecurity of people, crops, and animals. A primary challenge in all these areas is mapping genotype to phenotype through the complex web of evolutionary and mechanistic relationships among biomolecules, cells, organisms, and their communities in the context of the abiotic environments in which they live.

The number of identified genes and genomes is growing at an astounding rate, and the relationships among them scale combinatorially. Bacteria and their viruses alone account for approximately 10^{30} and 10^{31} species on Earth, respectively, with 10^5 bacterial species per gram of soil. BER resources are tasked with enabling their scientific communities to optimally access, analyze, and derive actionable predictions from biological systems data, which span sequencing, biochemical, and other biological features as well as a plethora of corresponding environmental measurements.

Integration of mechanistic and AI models and increasing efforts to combine genotype-based inferences with phenotypic screening are transforming the landscape of predictive modeling of biological systems. However, integration efforts are significantly challenged by the breadth of biological diversity at any given scale and the dependence of processes at a given scale on variables defined at different scales. While striving for a unified, quantitative description of living systems for the purpose of environmental sustainability and energy resilience, biological systems science still consists of a mosaic of subdisciplines with overlapping but distinct cultures, data infrastructures, and analytical tools. Data types are often similar across these different subdisciplines, but the different ways these data are annotated, stored, and made available create barriers that limit progress in this crucial endeavor.

2.2.1 What Research Could a Unified Data Infrastructure Enable?

A BER priority involves implementing a coordinated scientific strategy to support the development of multiscale biological system models that provide insights into molecular flows from single organisms to the planetary scale with the goal of ensuring energy and environmental stability and sustainability. Quantitative, predictive capabilities for biological processes are pursued as a key tool that will help inform policies and plans, ensuring the security and resilience of the nation's critical infrastructure and natural resources.

A barrier to addressing this challenge is the multidisciplinary and multiscale nature of the problem. Ecosystem-level processes may be directly affected by specific proteins or protein networks, specific microbes and their interactions, and complex microbiomes and their environmental feedback loops. Emergent properties at one level (e.g., division of labor across microbes) may be crucial for informing larger-scale dynamics (e.g., global energy flows), which in turn may constrain parameters for more detailed subsystem models. Moreover, complex feedback loops create cross-system couplings that cannot be disentangled without careful simultaneous consideration of different processes. For example, despite the availability of large microbial ecosystem datasets, it is unclear how microbial genomes and rapidly changing environmental parameters determine microbiome structures and, conversely, how the metabolism in these communities further influences external environments.

These challenges are not purely conceptual: To cope with a warming and rapidly changing planet, predictive multiscale frameworks that integrate different sources of biological information are urgently needed to learn how to efficiently grow plants and harvest valuable bioproducts under a varying climate (see Box 2.4, p. 15). Furthermore, integration of biological data across scales can help build the capability to engineer plants, microbes, and ecosystems capable of absorbing and retaining carbon. The need to develop and maintain forests, food crops, and biofuel crops at an increasing pace requires data integration and hybrid modeling approaches that can lead to optimized agricultural and management strategies. In addition to understanding how microbes, plants, and environments interact with each other, practical strategies are needed to design and deploy resilient energy crops that are efficient sources of bioenergy and biomaterials. This will require linking plant and microbial community genetic variation to environment variables and a deeper understanding of microbiome function, including how microbial processes impact terrestrial carbon and biogeochemical cycles.

Data relevant for advancing these goals are growing rapidly, both in coverage (e.g., number of sites and samples harvested, including through remote sensing and imaging technologies) and depth (e.g., amount of data collected per sample, including, most notably, single-cell datasets). Single-cell data provide greater resolution on the genetic and phenotypic variation in populations. This variation, in turn, could provide important insights into the capacity of organisms and ecosystems to respond to changing environments and to evolve toward newly adapted states in novel environments. The large diversity of scales represented in these datasets, if explored through the right quantitative tools, has the chance to provide unprecedented insight.

The quantitative tools to translate data into actionable predictions are rapidly expanding as well. In particular, the AI/ML renaissance, together with broad recognition that integration of mechanistic approaches with AI will be a key element of future predictive models, offers a unique opportunity for major leaps in multiscale biological systems relevant to BER. In parallel to the global scientific community's excitement about AI's enormous potential, caution and concerns permeate the field, as the trustworthiness and verifiability of AI models remain major sources of uncertainty and debate. Efforts to overcome this challenge will greatly benefit from broadly accessible, standardized, and diverse datasets. Increasing reliance on AI/ML requires a commensurate increase in quality standards and verification mechanisms for diverse datasets. A future comprehensive hybrid approach to predictive biology across scales could benefit from the creation of a comprehensive network representation of biological systems, where interactions, interdependencies, and correlations are all simultaneously accessible for further analysis, integration, and computation.

Box 2.4 Plants for Sustainability

Outcomes: Efficiently grow plants for production of biofuels and bioproducts under a varying climate.

Challenges: Plants, microbes, and environment/climate constitute a triad of intertwined systems associated with datasets often treated as separate repositories with different standards and cultures.

Unified Data Infrastructure Improvements: (1) Curate plant and microbial omics datasets and integrate them with field-collected data. (2) Make climate data accessible to plant and microbe researchers. (3) To build links across different research groups, support for the long-term data unification efforts that historically have been limited to short-term collaborations in specific crops. (4) Develop multiscale modeling approaches to connect genetic variation to production in future scenarios.

Future Benefits: Plant and microbe researchers will be able to access climate data and overcome the limitations of short-term collaborations in specific crops. Additionally, appropriate data infrastructure will facilitate the development of multiscale modeling approaches to connect genetic variation with production in future scenarios, aiming to efficiently grow plants for bioproducts and biofuels under varying climates.

2.2.2 Gaps: What Limits Progress?

Workshop participants discussed a need to effectively use multidisciplinary datasets to gain insight into plant and microbial systems biology as well as community interactions that affect larger-scale processes (up to the mesoscale). The biological data include sequencing (e.g., genomics, amplicon, shotgun metagenomics, and transcriptomics), proteomics, metabolomics, structural biology (e.g., enzymes, proteins, and complexes), and imaging data. While many biological science disciplines were early adopters of data adhering to the principles of openness and findability, accessibility, interoperability, and reusability (FAIR) (e.g., genomic data and GenBank), similar efforts within some of these disciplines lag. In several cases, finding, combining, and using multidisciplinary datasets are technically feasible, but the level of effort required to do so is often prohibitive. Thus, existing data hurdles slow and sometimes halt progress despite a recognition of the importance of grand challenge questions requiring integration of diverse data and researchers' desire to pursue these questions (see Box 2.5, this page). Highly manual and labor-intensive approaches carry a steep cost and do not support the iterative model development, testing, and validation approach encouraged across BER.

Data Discovery

Researchers must visit multiple data resources to find and access—typically through manual downloads—the data that they need. This is not a problem for data within their immediate area of expertise and stored in formats with which they are familiar. However, when stretching into new or adjacent subject areas, data discovery limitations can pose a challenge, according to feedback from both workshop participants and RFI respondents (see Fig. 2.3, p. 17). An immediate problem is the discovery of the resource itself. Many DOE user facilities (e.g., JGI and EMSL) or other data resources store and serve their own data, with some notable exceptions such as synchrotron data, which currently lack any such repository. Finding and gaining access to datasets from other agencies' data resources (e.g., NSF's NEON database) can pose an even bigger hurdle. Workshop participants expressed a need for additional training and easier-to-use data formats to assist them with finding data and being confident that they understand the data sufficiently to use for their own research questions.

Box 2.5 Microbial Data Integration

Outcomes: Understand how microbial genomes and the environment determine ecosystem-level structure and functioning.

Challenges: (1) Incompatibility of data types that together could provide extremely valuable insight but are currently siloed into separate formats and processing pipelines; a prominent example is the dichotomy of 16S rRNA amplicon data and shotgun metagenomic sequencing data. (2) Presence of nonstandard data scattered across studies. For example, several studies include microbial co-occurrence networks and microbial interaction networks obtained using different approaches. The lack of a standard for encoding these data limits the capacity for comparison and integration. (3) A lack of environmental metadata. Many studies involve microbial datasets obtained under specific environmental conditions. Details and formatting of this crucial environmental information are highly variable and often left to the discretion of individual researchers.

Unified Data Infrastructure Improvements: (1) A framework for data integration across various scales, from genes to communities. (2) Standardized metadata encoding and sharing.

Future Benefits: Improved understanding of how microbial genomes interact with the environment to influence ecosystem structure and functioning.

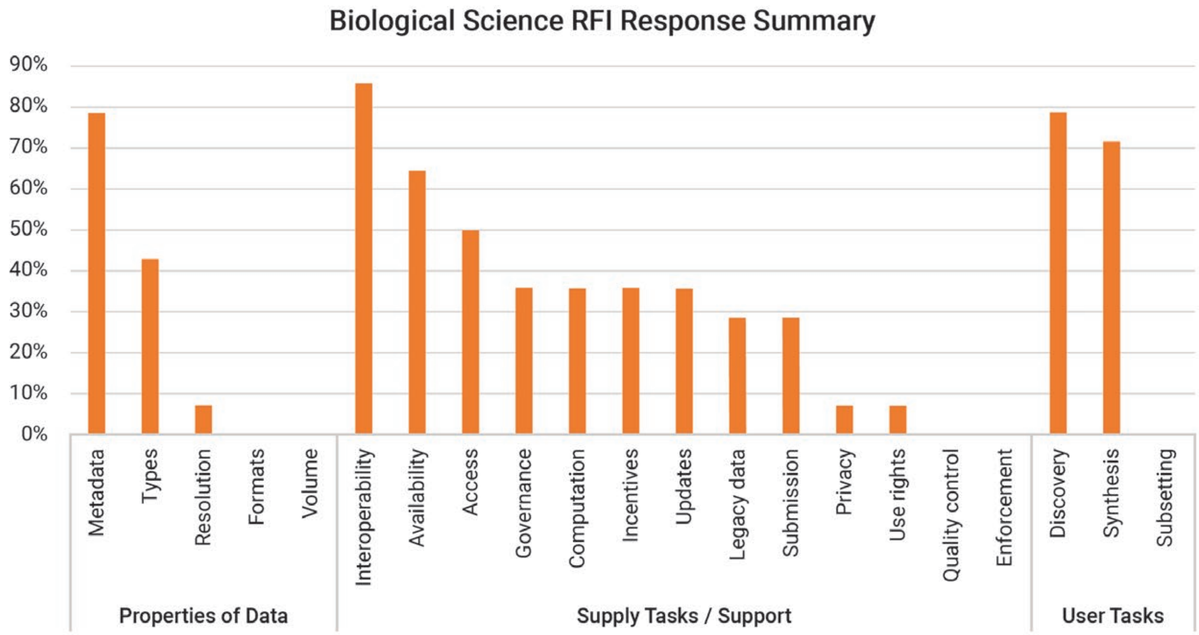


Fig. 2.3. Summary of Data Unification Challenges in Biological Science. These challenges were identified in written responses to a Request For Information; the percentage of RFI responses that noted each challenge is shown.

Combining Datasets

To combine datasets, the quality and comparability of each dataset from each source must be assessed. The methods to accomplish this may vary across data resources. In some cases, this assessment is possible prior to initial data transfer or download; in many cases, though, it must happen after the fact. Researchers then must develop their own system for keeping track of which datasets are useful within their downloaded files.

Workshop participants shared experiences where, in some instances, they wanted to combine data from the same samples that were stored in different data facilities. Often, these facilities and resources assign a unique ID to sample data. Thus, subsamples from a single sample that were sent to multiple facilities may be assigned different IDs with no obvious link among them. Even for the submitting researcher, reassembling the complete dataset from these multiple IDs is challenging, but it may prove altogether intractable for other researchers wishing to use these data as context or as part of a larger study. The benefits of BER's Facilities Integrating Collaborations for User Science (FICUS) program, where users can submit proposals to use multiple facilities within the same project, were highlighted as a useful approach. Critical limitations were noted, however, in that FICUS awards support only the original researchers; they do not inherently support data reuse by unaffiliated researchers.

When researchers combine data from different sets of samples or projects, major challenges include a lack of standardization in data formats, metadata formats and content, and the inability to translate between datasets using different ontologies. Lack of specificity in data descriptions also is a concern, as many similar measurements are not directly comparable. For sequencing data in particular, workflows and data processing approaches need to be captured in a standardized way to verify whether the data can be combined.

Using Data

After appropriate datasets have been located and harmonized, they must be transferred into the same shared computing space for analysis or modeling efforts. Although workshop participants did not express concerns about data file transfers, they noted challenges associated with the availability of compute capacity for performing analyses. Participants also felt they would benefit from additional training and examples of both curated datasets and workflows with expert commentary and discussion of best practices.

2.2.3 Discussion and Takeaways

The status quo provides little to no incentive for existing resources to collaborate or coordinate efforts to make data findable, accessible, interoperable, and reusable (FAIR) across resources. Even for BER data resources (e.g., JGI, EMSL, ESS-DIVE, ESGF, and ARM), significant effort is needed to determine how to connect them, identify equivalent terms, use common sample IDs, and take many other steps to enable researchers to seamlessly harmonize data. Ultimately, during funding renewal cycles, each resource must advocate for their individual, rather than joint, contributions to and impact on the scientific community.

Linking Data Through Standardization, Training, and Incentives

A critical first step in fully leveraging today's wealth of biological data is the ability to link DOE user facilities and major data repositories through standardized APIs and IDs. In response to broad community awareness of the need for standardized IDs, an *ad hoc* Samples Interoperability Working group has been developing standards for sample IDs, metadata assignments, and formatting across BER user facilities. However, as a largely unfunded effort, the working group's progress has been slow. More generally, workshop participants feel that increased use of standard ontologies across the community is necessary.

A lack of effectively enforced data and metadata standards greatly challenges efforts to seamlessly harmonize data from different sources. Although many data and metadata standards exist and are under development, the research community has not widely adopted them. To speed adoption, additional training and resources could help lower barriers for researchers not yet familiar with the standards.

Also, rather than requiring that collected data and metadata adhere to a standard that could be further refined over time, an alternative is to incentivize high-quality data management practices within the research cycle itself. One such approach, which follows the current paradigm for peer-reviewed papers, is to provide credit and citable digital object identifiers (DOIs) for data. This would validate and highlight data that adhere to standards and enable data producers to receive credit for data reuse. Another non-mutually exclusive option is to make analysis tools, models, and computing capabilities available to researchers who contribute data to a common system that requires adherence to community-developed standards at upload. This method incentivizes the researcher as an integral part of the research process, allowing for self-interest in research progress to motivate and increase compliance with data and metadata standards.

Improving Data Discovery, Findability, and Search

Even if all biological data were standardized and annotated for enhanced usability and interoperability, a major hurdle for data integration is the lack of a centralized catalog for discovery and findability. Such a catalog would ensure researchers are aware that such datasets exist and facilitate dataset discovery based on different criteria, even (and especially) beyond an individual's area of expertise. An effective, scalable, and federated search engine for both data and software applications is a major priority.

Many effective data facilities are available to the research community, but determining where specific data are located for use in particular applications can be difficult, as can knowing where to deposit data once collected. Similar to geographical maps of data developed for microbiome datasets, for example, vertical maps are needed to connect datasets across multiple levels, such as fluxes at the cellular and ecosystem levels.

Some data types (e.g., imaging data) are shared within specialized communities but are largely unknown outside of BER's current data ecosystem. Other data types are available in some BER databases but not integrated with others (e.g., structural data are not available in KBase). Diverse datasets thus need to be searchable and made available with enough context to enable researchers to easily determine whether they are useful to address a new question.

Developing an Infrastructure with Flexibility

Usability of diverse datasets by many different researchers requires the sharing of common workflows, analysis tools, and predictive mathematical models in association with their corresponding datasets. Importantly, the mode of data interaction strongly depends on the approach taken and should be flexible enough to accommodate different types of inferences, predictions, and testing. For example, some approaches are strongly data-driven and can identify patterns irrespective of underlying assumptions about phenomenological laws; others are deeply mechanistic and may require systems biology modeling scaffolds. It is particularly important that the next data infrastructure is flexible enough to facilitate integration between AI and mechanistic models.

2.3 Crosscutting Science

Several common gaps and opportunities emerged across the environmental and biological sciences RFI responses and workshop breakouts (see Fig. 2.4, this page). In particular, three crosscutting themes apply across much of BER's science domain and thus constitute broad areas of potentially high impact: (1) leveraging decades of environmental and biological site data; (2) developing and testing multiscale models; and (3) making BER environmental and biological data broadly accessible, inclusive, and usable.

Leveraging Decades of Environmental and Biological Site Data. The goal of this first crosscutting challenge is predictive understanding of biological and environmental systems under realistic field conditions at multiple scales (see Box 2.6, p. 20). Across environmental and biological sciences, research

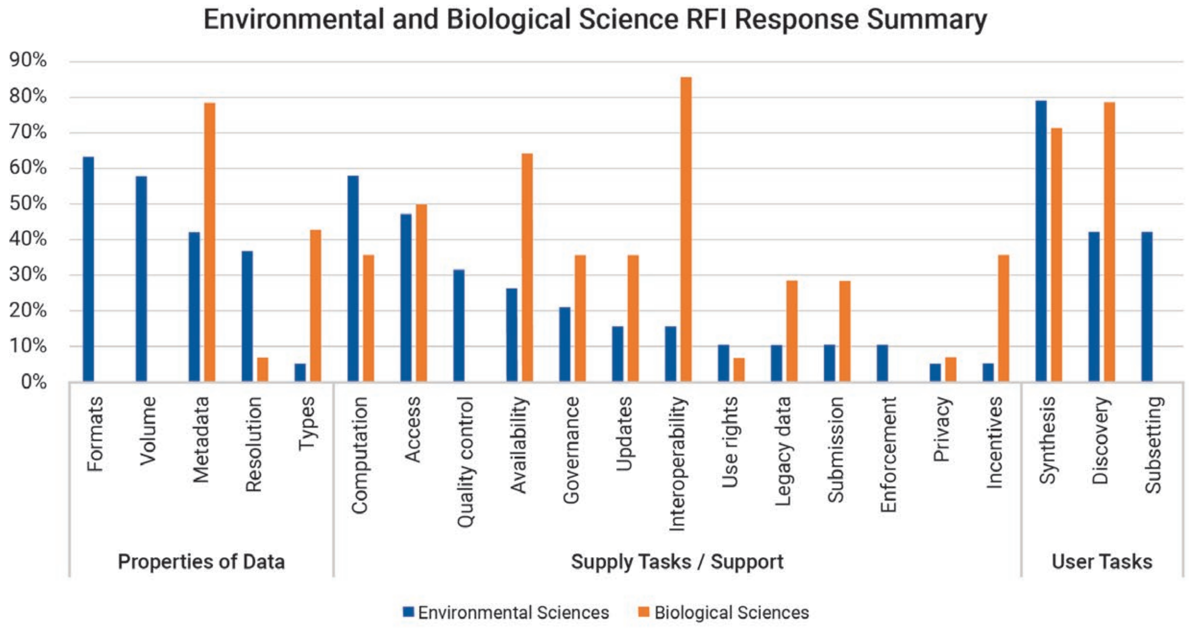


Fig. 2.4. Summary of Combined RFI Responses of the Biological and Environmental Science Communities. Data categories appear in descending order of importance for the environmental science community.

Box 2.6 Using Outputs and Insights from Climate Models to Improve Biofuel Crop Design

Outcomes: Incorporate localized future climate prediction models into the development and deployment of specific genotypes for biofuel grasses and tree crops.

Challenges: (1) Efficient biofuel feedstocks are long-lived perennials that will need to survive 10 to 15 years once planted in a location, but breeding, selection, and engineering of these plants are for the current local environment. Deployment will likely be as replicates or narrow germplasm to meet production needs. (2) Current design itself is lacking sufficient linkage of genotype (gene-level variation) to environmental impact data, and new genotypes will take many years to develop and prepare for deployment. (3) Current collected data for multisite testing of genotypes are not accessible for community development and can only be curated by a small number of experts.

Unified Data Infrastructure Improvements: Storing and streamlining access with APIs to curated datasets of genotypes, gene expression, field phenotypes, and field experimental datasets would enable advanced algorithm development, including AI approaches, to connect genotype-level variation with field performance under varying climatic conditions. Co-locating this data with access to curated, geographic-localized climate models would enable the development of predictive performance models for existing genotypes and direct genotype development for focused areas of deployment under specific future climate scenarios.

Future Benefits: Biofuel deployment will be a continuous development process. Even as the first successful genotypes are put in the ground, the next version will need to be in the scale-up phase for deployment. These genotypes will need to be continually tested across multiple climatic zones and the data integrated to guide the next round of selections and targets in the breeding populations. Many groups will be needed to produce and test these crops across many locations and potential applications. A unified data infrastructure would underpin this process and computationally accelerate the development of new genotypes to adapt to rapidly changing climatic conditions.

groups commonly engage in laborious, repetitive manipulation of diverse datasets when synthesizing, or even simply using, field data. Such effort is necessary because field data tend to be heterogeneous and emerge from disciplines that have few, if any, expectations or culture of data sharing. Consequently, research groups are forced to re-invent tools that could be universally adopted. Compounding the problem, there are few incentives for data integration because of diverse agency data sources and objectives.

Developing and Testing Multiscale Models. A particularly difficult aspect of this second challenge is incorporating environment-dependent prognostic biological system parameters (Todd-Brown et al. 2022) in high-resolution regional and Earth system predictive models (see Box 2.7, p. 21). This work typically requires specialized knowledge to access, handle, and interpret the disparate locations, ontologies, and formats of biological and environmental data that span a wide range of spatial and temporal scales (Todd-Brown et al. 2022). These limitations slow model development, iterative testing, and falsification of alternative model structures (i.e., the entire model-experimental process through which such science advances).

Making BER Environmental and Biological Data Broadly Accessible, Inclusive, and Usable. This third challenge is a relatively new priority for many scientists and program managers, especially as it pertains to working with communities, and the incentives, processes, and timelines for making data available and reusable remain unclear or weak. Moreover, few opportunities exist for data users and decision-makers to engage directly with each other, and there is a general lack of knowledge about how to access datasets and models to test decisions or usability.

In summary, these crosscutting opportunities represent areas in which BER can have significant impact, particularly in understanding and manipulating environmental and biological systems to meet DOE goals.

Addressing these challenges would require DOE to (1) encourage shareable, coordinated data collection with standards for field data; (2) develop curated, standardized, and open datasets that can address multiscale modeling; and (3) focus on making field, environmental, variation, and climate data accessible to support diversity, equity, and inclusion goals.

Box 2.7 Improving Prediction of Coastal Disturbance Impacts on Human Communities and Ecosystems

Outcomes: Seamless ingestion of diverse, multiscale datasets for a generational leap in coastal model prediction of the impact of rising sea levels and storms.

Challenges: Land-ocean interfaces exert disproportionate effects compared to their area. For example, coastal ecosystems account for <0.1% of the ocean surface but sequester ~50% of all carbon while also emitting substantial methane. These processes, and their impact on the human communities that cluster along coastlines, are difficult to model because they cross a wide range of spatial scales, with small-scale effects propagating to higher scales in a dynamic, spatially complex, and widely dispersed manner. A central challenge for coastal models is to leverage current knowledge of fundamental processes and ecosystem state changes into a computational framework that can robustly predict the impact of rising sea levels and increasing storms. Such models are currently data-limited, however, making performance improvements difficult.

Unified Data Infrastructure Improvements: A computational and data infrastructure that provides cross-scale capability, linking point observations (~1 m^2) to flux tower measurements (~1 km^2). This capability would enable turnkey data ingestion by the models used and pioneered by DOE projects. Such projects include Coastal Observations, Mechanisms, and Predictions Across Systems and Scales (COMPASS; compass.pnnl.gov), which focuses on coastal biogeochemistry, and Integrated Coastal Modeling (iCOM; icom.pnnl.gov), which integrates human and natural coastal system behaviors. This envisioned infrastructure would allow rapid model development and testing (parameterization, benchmarking, and prediction) by leveraging the broad spectrum of data collected at different spatial scales by the diverse science networks operating in U.S. coastal regions.

Future Benefits: Over 128 million people in the United States (roughly 40% of the population) live in coastal counties. Hurricanes have accounted for over $1.3 trillion in damages and more than 6,000 deaths since 1980. Improved modeling of coastal systems is necessary to adequately assess risk to and plan for community and energy infrastructure resilience in the face of the persistent pressures on coastal zones from sea-level rise and increasing frequency and severity of storm events.

3 Workforce Development, Inclusion, and Accessibility

This chapter discusses current barriers to data findability, accessibility, interoperability, and reusability (FAIR) and equitable access to BER science and how a unified data infrastructure may help in overcoming these barriers.

3.1 Improving FAIR Practices

3.1.1 Current Barriers

Workshop participants underscored the importance of FAIR data principles as tenets of a unified data infrastructure. However, they noted several challenges with existing implementation of FAIR data in the greater scientific community, including the *ad hoc* development of metadata and FAIR standards and the need for better governance of the data management and usage process across organizations and services. Improved governance and broader community contribution to the development of standards would facilitate greater recognition of the importance of using such standards, better scoping of individual standards for metadata, and more uniform adoption time.

Further, when considering FAIR standards compliance for a new or updated resource, the "findability" and "accessibility" principles are often viewed through the lens of the resource creators and their known community of users. Without an effort toward broader community awareness, this view may leave out a significant potential audience, particularly those without the connections to participate as first-hand collaborators. One way to alleviate this disconnect is by developing universal cataloging and common vocabularies or mappings between communities. Similarly, "interoperable" has often come to mean interoperable with the tools and users already familiar to a resource developer. The broader research community could benefit from stronger interoperability standards beginning with straightforward cases like common units.

It is not enough for data, metadata, and infrastructure to adhere to domain-specific standards and FAIR principles. Integration across scales requires (1) thinking beyond FAIR toward well-organized and comparable data; (2) creating global catalogs enabling cross-search and discoverability; and (3) embracing federation and automation to support standards implementations, annotation, provenance, and updates. These recommendations also raise questions about governance, equity, and accountability. Standards, policies, and catalogs must consider and respect the needs and circumstances of users and stakeholders, both large and small. Those developing these standards should actively invite input and feedback from as broad a community of potential users as possible.

Another challenge lies in addressing different metadata and FAIR data standards that various research communities may already have adopted for the same data based on their own specific needs. Where possible, developing automated or semiautomated conversion tools will be important to enable the flexible integration of these data across communities with different existing data and metadata standards.

In addition to the challenges of working with data itself, a significant learning curve for many researchers lies in data science and the ability to effectively access and use datasets. While some progress is possible through mandates and requirements, previous efforts have shown that these may not yield optimal results. For example, many issues with metadata submissions might be expected from researchers who are inadequately trained, do not understand the benefits of such data efforts, and submit metadata under time pressures associated with publications or progress reports.

Ensuring that a unified data infrastructure brings in high-quality data requires appropriate incentives. For small data producers (e.g., independent university researchers versus data facilities), these incentives need to be relatively larger because of the higher burden of effort for these producers and lower chance for effective workflow or automated solutions.

3.1.2 Recommendations

To address the barriers slowing progress on FAIR data and metadata practices, the subcommittee identified four areas of need, including:

1. A governance or guidance structure and support for community-driven development of targeted, domain-specific standards (e.g., Genomic Standards Consortium's env development packages).
2. Adequate incentives to ensure high-quality data are captured by the system. Incentives should include workflows or advanced analysis tools so that data submissions are integrated as part of the research process rather than completed after the fact.
3. Direct funding support, training, and tools to streamline and lower the burden of data management–related tasks for researchers, especially those from smaller projects.
4. Improved interoperability of data from multiple environmental fields and their integration with biological datasets to create new knowledge links.

3.2 Moving Beyond FAIR to Equitable

Many research communities have moved toward more open data and open science. This trend has potential to support an increasingly democratized and, ideally, equitable approach to science. Although making data and tools available increases equality of opportunity, this is not sufficient to ensure equity.

There are at least three constraints surrounding equity in data access and use: (1) access to a network or cohort of users that can guide the process and lower the entry barrier, (2) computational resources to work with the data and models, and (3) lack of incentives or rewards for users to invest in data curation and metadata creation over other opportunities for career advancement.

In an open science world, awareness of data, analytical tools, and models is prerequisite to their findability, accessibility, and reuse across research domains. Researchers currently well integrated into existing (and sometimes funded) research networks tend to have first knowledge of data and tool releases, and thus they are able to benefit more from new developments. Those who are not connected through these networks must put in considerable effort to learn about or find them. Even with these efforts, the delay may cause them to fall even further behind in leveraging the resources in their work.

Research network connectivity does not affect everyone equally. Researchers from underrepresented groups, from institutions where research is a requirement but not a primary focus, or those who are otherwise underfunded are less able to connect with and participate in these networks.

At many historically black colleges and universities (HBCUs), minority-serving institutions (MSIs), and community colleges, student research engagement efforts must recognize that students often support themselves or family members through other employment. Paired with opportunities that improve students' employment potential beyond scientific research, improvements to data infrastructure can reduce barriers and provide incentives for students to engage with data tools and activities. The availability of training in a unified data infrastructure also can enable the transference of skills and knowledge to workplaces beyond a researcher's laboratory.

In addition, campus computer centers and supercomputers focus on compute power and have limited staff available to assist with data analysis tasks. Access to common BER entry points for guidance on running core informatics tools and models would make data accessible to a more diverse group of users. Whether through instructional tools or expert assistance, this guidance could make data and models easier to use for researchers lacking expertise or previous exposure to these resources. This, in turn, could create a surge in researchers' ability to use the data and models in fields where they are not widely leveraged yet.

3.2.1 Supporting Community Development and Inclusive Networks

The success of any data infrastructure project, especially open projects that stand to benefit from outside contributions, depends on a productive community of users. A platform's success is more effectively achieved by engaging diverse user and contributor communities. Equitable community development does not happen spontaneously; it requires well-designed and sustained effort.

This effort must begin from the outset of project design. The first step in community development is the need to demonstrate both good intent and genuine interest. Workshop participants stressed the need for new and currently underserved community members to have a seat at the table early in the design phase of the project, both to ensure their voices are heard and to demonstrate the project's values. This could be accomplished by holding workshops and open sessions to listen to different communities' needs and understand their varied challenges. Such sessions would begin to build trust and provide the project with broader insights and input for planned development.

A unified data infrastructure has the potential to connect members across existing communities and networks. As these communities combine, recognizing any existing imbalances in scientific networks and systematizing efforts to address them will be important. This can be accomplished by creating effective network analysis and mapping, documentation, and readily available training designed to meet people where they are in terms of scientific background, experience, and culture. These documentation and training efforts will be critical to increase the ability of the user community to effectively address scientific questions that span current boundaries (e.g., those of scale or scientific domain); they can also serve to offset current inequities in scientific networks and level the playing field.

The broader BER community would benefit from the development of best practices for project teams that foster a welcoming environment that draws and sustains participation from diverse communities, including nontraditional new users.

Relatedly, workshop participants recommended creating an advisory board that includes members of underserved research communities. A diverse advisory board, including members from all career stages and different institutional types, is critical to the goal of identifying and meeting the needs of broad user and contributor communities of the data infrastructure. More specifically, the board should include faculty and scientists from HBCUs and MSIs, which would further solidify the project's commitment to equity.

3.2.2 Improving Engagement with Researchers from Underfunded Institutions

Workshop participants emphasized the primacy of mentoring, training, and networking to increase the engagement of researchers from underfunded institutions with the unified data infrastructure as well as existing data infrastructure projects. Specifically, the importance of human connection was emphasized through recommendations to directly reach out to schools by organizing events, conducting targeted recruitment for internships, in-person workshops and training, and networking opportunities. In addition to their value in increasing awareness of available data and tools, these activities are prime opportunities for BER researchers to expand human connections and develop peer mentoring relationships with faculty. These relationships can help researchers better identify training or internship opportunities as well as any existing impediments that could prevent use of the unified data infrastructure itself.

Building Connections Between Senior Scientists and Early Career Researchers

Participants also expressed a need to engage both emerging and senior scholars to develop connections between both groups. Early engagement activities could include researchers reaching out to students to discuss what it means to be a scientist and teaching them how to interact with data and the instruments that generate them (e.g., Girls Who Code and Pacific Northwest National Laboratory's Pathway Summer Schools program). Suggested activities for engagement with senior scientists included intentional development of interpersonal networks and long-term mentorship programs.

Raising Awareness of Data, Tools, and Projects Through Targeted Outreach

Workshop participants also discussed the importance of recognizing the difference between equality and equity, along with the accompanying need to provide resources and opportunities necessary for underserved research communities to have equitable access to data and tools. This requires a deep understanding of the differences in circumstances, an understanding that can be gained through relationships developed by the outreach and engagement activities described above. In many cases, researchers from underfunded institutions are disadvantaged by less extensive networks, resulting in a decreased awareness of available data, tools, and projects from the national laboratories. Additional targeted outreach is necessary to address this disparity. Furthermore, a centralized marketplace that provides information on ongoing projects, as well as available data and tools, is not only necessary and helpful but also a means of contact with researchers associated with those projects. Such a resource is critical to ensuring engagement. Parallel targeted funding opportunities that support faculty research and development, such as the Reaching a New Energy Sciences Workforce (RENEW) initiative, is also seen as critical to achieving an equal outcome.

BER has been conducting targeted outreach and actively seeking partnerships with HBCUs and MSIs. This strategic engagement opportunity by BER programs and partners is constrained by a lack of adequate technological access at underfunded institutions that inhibits their ability to meaningfully participate in and contribute to BER activities. The availability of a unified data infrastructure can bolster access for communities currently challenged by limited access to high-performance computing infrastructure. This access, in turn, will lower the barrier for participation.

Expanding DEI Opportunities

A final and critical recommendation is increased investment, both in time and resources, in creating more opportunities for BER researchers to learn about diversity, equity, inclusion (DEI). Such initiatives would advance BER's ability to relate to researchers at HBCUs and MSIs. The development of peer mentoring relationships between faculty and national laboratory scientists is mutually beneficial and vital for successful engagement. Workshop participants also highlighted the need to create local and regional project clusters that invest in and build partnerships with HBCUs, MSIs, other universities, and local communities. New research avenues are increasingly incorporating a greater societal impact component that considers both new data types and a wider range of potential data users, such as decision-makers and community members who are nonexperts.

Recent BER initiatives such as the Urban Integrated Field Laboratories (see Box 3.1, p. 27) directly engage and impact stakeholders and community partners. For such projects to be effective, community partners, faculty, and scientists with diverse expertise and experiences need to be active project decision-makers, not just beneficiaries of project outcomes. The active engagement process creates avenues to involve in BER activities students and early career scientists with varied backgrounds, including social science, climate applications, and noncomputational or nonmathematical expertise.

Many community partners outside of universities and government agencies may not have computational backgrounds or the requisite computing power to address their problem/solution approach. In such cases, providing tools that can be used with limited processing is important for expanding data and tool usage. For example, BER data needs to be integrated with other datasets (e.g., from NASA, the National Oceanic and Atmospheric Administration, the U.S. Department of Agriculture, and international centers) for environmental and community-facing applications. Linking field data from BER's Atmospheric Radiation Measurement user facility with climate model output that can enable comparisons would ease and broaden the use of such data studies. This integration would allow BER data, especially datasets with strong societal linkages, to become more accessible, inclusive, and usable by the research community, as well as more readily integrated with student and project research plans.

Box 3.1 Urban Integrated Field Laboratories

The Urban Integrated Field Laboratories (UIFLs), established through grants awarded in 2022 by BER, present unique case studies for a unified data infrastructure. Although each of the four UIFLs (ess.science.energy.gov/urban-ifls/) has a region-specific focus, commonly shared research themes include understanding climate, air, and water patterns and challenges related to urban environments in order to inform the development and actualization of resilient community solutions. Through established and expanded networks, the UIFLs bring together their surrounding communities and stakeholders to co-design the science and solutions to promote climate resiliency. Such profound yet broad undertakings generate large amounts of data and necessitate not only the standardization of data streams but innovation in the ways in which those streams may be curated and shared.

Types of Data and Uncertainties

To promote resiliency, teams must acquire, combine, and make actionable basic science data. The data that exist and will be gathered by the UIFLs are at different scales (e.g., human/household scale to community scale to regional scale). These data arise from disparate fields and theme areas including climate/heat, air quality, hydrology, planning, and social factors related to community vulnerability.

Moreover, different datasets and outcomes may have different levels of uncertainty associated with them due to the methods by which the data were gathered. Disparate methods include (1) curation and standardization of data from persistent observation networks (for science measurements) and environmental justice platforms (for social factors); (2) generation of new data from mobile and stationary environmental monitoring platforms; (3) modeling data; and (4) data from community co-participants and co-designers. Some of the uncertainties may be relatively fixed and easily quantifiable using traditional single-discipline data (e.g., the measured concentrations of pollutants using low-cost sensors versus the concentrations measured from traditional research-grade instruments). Other uncertainties may be more variable and involve the coupling of science measurements and social determinants. As a specific example, the uncertainty in the number of individuals impacted by a heat mitigation solution may involve both temperature data and community data if there is an unhoused population that moves due to social factors.

Data Uses

UIFL research is uniquely positioned to impact both the UIFL's local population and the broader scientific community. The approaches to coupling and utilizing disparate datasets must be flexible to accommodate different end users. Some end users of UIFL data may be community members that wish to evaluate tradeoffs between potential solutions given varying scenarios. Thus, providing facile methods of accessing and visualizing data and changing scenarios is essential. Other end users may be scientists that use computer-based techniques (e.g., machine learning) to bring together disparate pieces of either pure science information or coupled science and social information to train models and enable predictions. In each case, enabling facile rescaling of output data and formats to address different constituents is key. Moreover, sensitivity to protecting privacy must be factored into how the data are shared, especially when using data at the individual human or household scale.

3.2.3 Supporting Workforce Development

BER workforce infrastructure is affected by the lack of diversity in science, technology, engineering, and math (STEM) programs across the United States and globally, and this diversity progressively decreases when moving from the undergraduate, graduate, and postdoctoral levels. BER has an important and exciting opportunity to engage early career scientists in its research portfolio, which tackles societal problems that require integrating hyperlocal social vulnerability, ecology, and large-scale climate change issues into analyses. The intersection of such critical problems and students' high interest in them is an opportunity to engage students beginning at the undergraduate level and continuing throughout the workforce development pipeline.

Unified Data Infrastructure's Role in Building Collaborations

Creating a unified data infrastructure framework is one way to mitigate or lower barriers for engaging a diverse workforce. Another is embracing reflexive, coproduction engagement for problem solving, which ultimately will increase awareness of the research infrastructure and engage diverse partners and early career scientists in BER science.

Data creation, use, reuse, and analysis are often the currency for collaboration across projects and teams. A unified data infrastructure and governance framework can help highlight and promote collaborations that benefit from sharing standardized data within the broader BER community. In turn, this will help researchers—especially students and early career scientists—recognize the potential value of developing datasets and collaborating with teams to advance their careers, thus infusing a new focus on research data management.

Expanding Training and Outreach

To facilitate these collaborations and engagement, workshop participants highlighted the value of increased training activities and improved outreach for existing activities. In particular, efforts should be made to develop training that is accessible to diverse communities, including modifications that support place-based inquiry. Workshop participants discussed the importance of adjusting training materials to consider the specific problems of interest to researchers from underserved communities. In other words, the burden of making the training relevant and meaningful lies with the trainer, not the trainee. Given this responsibility, and the range of adjustments that may be necessary to ensure the relevance of training, a train-the-trainer model (e.g., The Carpentries) was suggested. Participants highlighted the potential benefits of having early career scientists in this role, both for their own careers and the larger community. One existing model for this within BER data infrastructures is the educators' community within the DOE Systems Biology Knowledgebase (KBase) and its efforts to build a microbiome workforce development program.

Specific training topics essential for workforce development include standardized sample collection and processing, best practices for accessing and using open community-collected data, and support and training in data analysis and publishing. Workshop participants suggested demonstrating the overall benefits of a unified data infrastructure by showing how data collected by diverse communities at different institutions to address distinct questions are relevant to their individual needs and experience; such datasets could be used to address larger questions because they are standardized and contextualized to be reusable.

Capitalizing on Growing Interest in AI/ML

BER-centric problems are accessible and "understandable" by the diverse scientific community. Working with a unified data infrastructure provides access to data and promotes a more diverse and equitable scientific enterprise once barriers to both data and models are lowered through open access codes, guidance on their application, and identification of data resources and models. Recent reports document the emergent opportunity arising from current interest in using artificial intelligence (AI) and machine learning (ML) activities to engage diverse early career researchers and students. Topical areas like Earth system digital twins and the role of AI/ML predictive models in Earth system modeling appeal to the community and provide an avenue to engage and build a diverse workforce.

The National Science Foundation and collaborating agencies have launched National Artificial Intelligence Research Resource (NAIRR) pilots to provide access to advanced computing, datasets, models, software, training and user support to U.S.-based researchers and educators. Similarly, BER has an opportunity, in coordination with DOE's Advanced Scientific Computing Research program and other department activities, to create a platform to build the profile for such workforce engagement and development.

Overcoming Data Literacy Challenges

Data literacy is a barrier for the broader community not currently engaged with DOE data creation and analysis (see Box 3.2, this page). For many early career researchers, data standards remain unfamiliar, and more training and support are needed for improved

Box 3.2 DATA LITERACY

"Data literacy" is often considered to be the ability to explore, understand, and communicate with data in a meaningful way. However, achieving data literacy assumes that the data and tools used to work with data are robust, well documented, and readily available. Advancing data literacy within the scientific community will require an intentional, community-wide effort, especially by scientists, engineers, and data managers. Current efforts to advance data literacy are underway, including the 2023 Year of Open Science and various efforts by groups such as the Research Data Alliance, Earth Science Information Partners, and others. And while these efforts try to work across agencies, the depth of adoption within any one agency is still limited.

Standards and Conventions

Establishing and using data and metadata conventions and standards can enable increased data literacy. Standards such as the Climate and Forecast conventions (CF; cfconventions.org) are designed to make associated scientific datasets easier to process and analyze. Metadata standards define the vocabulary used to describe the data recorded in a dataset. When collections of data are well documented, datasets are easier to analyze because the structure and contents are easily determined, enabling researchers to focus on analyzing and understanding the information's significance. This is the core tenant of data literacy.

Role of AI

In the coming decade, achieving data literacy will undoubtedly shape scientific understanding and capabilities in numerous fields. In the near-term, a pivotal role is foreseen for artificial intelligence (AI), which is projected to not only change analytical approaches but also be instrumental in data management and data wrangling processes. As the research community navigates this era, the understanding of data literacy must evolve to include a deep integration with AI, especially when considering the impact of generative AI in scientific domains. This change involves more efficiently leveraging large, established models, such as foundation models, and equipping researchers with the know-how to adapt these models through fine-tuning, prompt engineering, and various tuning techniques. Moreover, mastering a systematic approach to verification and validation becomes imperative. The growing interest in digital twins and generative models, as well as topics related to explainable AI, are part of a rapidly growing domain.

Another aspect to consider is the impact of large language models (LLMs) on the data and research lifecycle. With their disruptive potential, LLMs can significantly modify existing workflows. An equally significant area of focus is AI ethics, which delves into not only the behavior of AI models but also the datasets used for their training, as well as their subsequent real-world applications. Enhancing data literacy by focusing on the systematic evaluation and validation of AI model outputs will become increasingly necessary.

adoption by a diverse community of researchers. This is especially true for scientists who have not previously participated in large, collaborative projects that require data exchange across many research groups. These researchers may be less familiar with the necessary details and steps required to carry out this exchange as well as the benefits of doing so.

Limited cross-domain training and vocabulary, along with mismatches between standards and metadata across the range of scales addressed by BER research, affect collaborative work across such scales. The challenge is notable for underfunded academic institutions but is also prevalent in large BER communities among R1 institutions not directly engaged in data creation, data access, or computational work. BER community researchers have heterogeneous and fragmented skillsets in data infrastructure and management. As such, many researchers, especially early career scientists and community members, are excluded because they do not have the direct expertise or resources to invest in dedicated data processing and analytics staff.

Many graduate students and early career scientists are engaged in hypothesis testing or process-scale studies. For many of these researchers, data management introduces a novel challenge. A unified data infrastructure framework could help them collect and organize their data. Presently, smaller project teams find the task of data standardization onerous, impractical, and disproportionate in terms of effort and reward (i.e., low returns on high efforts invested). Training scientists to help think about potential interoperability and uses for their data beyond the scope of their own objectives may enable broader collaborations. Opportunities to develop such data-driven interoperable analyses as a project supplement can help incentivize such efforts.

A uniform data infrastructure enabling community contribution of data, analysis workflows, and modeling tools, including AI modeling, allows for real-time, interactive sharing of research code and pipelines. Tracking individual contributions and the extent of their reuse could increase attribution for scientific work that has traditionally not been credited or cited, potentially allowing researchers to receive recognition for their contributions earlier in their careers. This outcome depends on shifting the research culture to equally value and reward these important contributions and consider them in career and promotion decisions. A unified data infrastructure can help drive this culture shift by supplying the necessary metrics and directly promoting the contributions of early career researchers. By serving as a central marketplace, a unified data infrastructure would also help these early career researchers build their professional networks.

3.2.4 Recommendations

To address the challenge of developing an equitable unified data infrastructure, the subcommittee identified eight necessary steps:

1. Target diverse stakeholders in outreach efforts during the initial design phase and sustain support for these efforts and for mentoring as data and tools come online. The Carpentries model of training regional or community-specific trainers is suggested as a best practice for this sustained outreach effort.

2. Prioritize, within the unified data infrastructure itself, a simplified user interface and single point of entry to BER data facilities. Implement a search functionality across multiple data repositories and a searchable catalog of tools and paths to finding relevant data for tool use.

3. Provide compute resources for data processing and tool use to support researchers from low-resource institutions.

4. Provide to researchers from a broad set of backgrounds accessible training that includes a mix of (a) publicly available and reproducible examples; (b) clear and detailed documentation; (c) web-accessible video tutorial content; (d) example reproducible workflows; and (e) workshops at scientific conferences and by request.

5. Develop data-centric DOE graduate and postdoctoral fellowships that encourage early career researchers to invest in BER data activities and become part of the broader BER ecosystem.

Increasing such fellowships within BER will elevate these opportunities into meaningful pathways. Moreover, workforce requirements for AI-driven future priorities necessitate the broadening of eligibility requirements for these fellowships to include all academically eligible STEM graduate students and recent graduates from U.S. universities, as is currently done for NASA educational programs.

6. Provide merged access to select datasets managed by other agencies. Climate datasets have become useful within the broader BER context and need to be made available for easy use by partnering with regional hubs and climate offices that translate information to end users. Strategic funding calls can also support such data use and guidance hubs.

7. Provide supplemental funding after an initial grant period to make project-generated data interoperable, accessible, and reusable.

8. Explore partnering with other agency programs that have experience with DEI initiatives and workforce development, with the added benefit of encouraging cross-agency collaborations.

4 Unified Data Infrastructure and Artificial Intelligence

This chapter summarizes barriers to scientific progress identified by the subcommittee that could be addressed by a unified data infrastructure across BER programs, projects, and facilities. Some of these barriers could be overcome by current BER capabilities available at BER supporting user facilities, for example. For the remaining barriers, the chapter discusses whether potential solutions require policy changes, implementation of known solutions, or research. Finally, recommendations are offered on progress that might be achieved in developing a unified data infrastructure in the 5-year time horizon specified in the charge.

4.1 Barriers to Progress

Scientific challenges in the biological and environmental sciences increasingly require the integration of multidisciplinary and multiscale datasets. However, specific challenges vary by community. In environmental science, data are usually available and accessible but exceedingly large, and several datasets often need to be integrated for research. In the biological sciences, datasets are generally more manageable but often sparse and incomplete, and not all data is accessible to a broader research community.

For both the environmental and biological communities, the size of datasets makes repeated downloads, transfers, and processing infeasible for most users but disproportionally affects those at institutions with fewer resources. Additional challenges emerge when trying to integrate data across modalities, studies, and projects. New imaging technologies also have rapidly increased data volumes and associated analysis challenges due to data size and analysis complexity (e.g., cryo-electron microscopy and cryo-electron tomography).

4.1.1 Working Practices for Data and Compute

Current working practices related to data and compute locality pose significant barriers to scientific progress. The practice of locally downloading all required datasets is unsustainable, given the increasing volume of data from instruments and computation. Also, a significant number of project-specific biological datasets are not shared or openly accessible, and data and software assets provided for sharing generally need more contextual information and quality controls to ensure their proper use in an experimental context. The additional effort required to enhance data reusability is a huge burden for researchers.

Furthermore, an increasing number of participants in BER-related research lack awareness of available datasets, workflows, tools, and frameworks, leading to an underutilization of existing capabilities and resources as well as a duplication of effort. This has resulted in a proliferation of tools and led to knowledge, data, and skill gaps regarding ever-changing software, tools, and frameworks. Gaps in data volumes and skills are particularly steep barriers to early career scientists and researchers from minority-serving institutions with fewer resources and capabilities.

4.1.2 Unified Search Capability

Scientific research today is increasingly complex and relies on many different sources of information. Fully leveraging this wealth of data requires a unified search capability for BER science. Researchers need to be able to search metadata and data across different archives and assess and access what they need from each archive in one single activity. Also essential, but currently lacking, are standard interfaces and tools to access data across diverse data repositories, a common clearinghouse for metadata, and consistent minimum metadata standards. Another key barrier is the inability to analyze data where they are stored. These challenges are further exacerbated by the need for persistent, globally unique identifiers for data across multiple systems and for commonly adopted conventions for variable names and units.

4.1.3 Data Standardization and Accessibility

Working with data from different sources is further hampered by a lack of standardization of data and metadata formats. This challenge arises within some BER science domains and when aiming to integrate data across domains and archives. Not all data created by BER research is easily accessible. Instead, efforts to locate data can often require personal contacts or in-depth community knowledge, and once located, the data are not easily accessed because no standardized protocols or APIs exist. These are key hurdles for all researchers but disproportionately affect early career scientists and those new to BER science.

More comprehensive metadata could encourage increased reuse of BER datasets and broaden their use beyond the communities that originally created them. However, the inability to track data reuse over time discourages those willing to share their results, as they lack the incentives and rewards for their efforts. Adopting data standards would make the tracking of data reuse tractable. Systems like Altmetrics or Web of Science can ingest standardized information to aggregate statistics on reuse, providing similar measures to an h-index to quantify a scientist's impact.

4.1.4 Data Integration

Integrating data from multiple sources into one coherent body of information for further analysis remains a predominantly manual task that includes, for example, harmonizing and translating metadata and resolving conflicts between metadata descriptions in different fields (e.g., the many different definitions of temperature). This is followed by even more time-consuming work to harmonize and integrate data formats and content. These data integration tasks present a significant barrier to working with data from different sources. While a fully automated solution for this challenge cannot be expected soon, tools are needed to lower this barrier.

4.1.5 Localized Solutions

The lack of a unified infrastructure for datasets and workflows results in partial and local solutions that do not support scientific research at the scale and level of sophistication needed by the BER research community. Furthermore, the subcommittee sees massive inter- and intra-agency barriers to cast temporal and geospatial maps, for example, and rapidly contextualize regional and local data (without having to download "everything"). There is no ability to create temporal data hubs where disparate datasets can be brought together to solve grand challenge problems in the BER mission space, such as those outlined in Ch. 2: Science Opportunities, p. 7.

4.1.6 Unified Data Storage and Processing Capabilities

Additional needs arise from the lack of unified data storage and processing capabilities. For example, a common platform is needed for observations (similar to the Climate Model Intercomparison Project and Earth System Grid Federation) and for experimental results (similar to the DOE Systems Biology Knowledgebase, or KBase). Also needed is readily expandable and scalable data storage capacity, similar to Amazon Web Services or Google, but without the worry of losing data if storage becomes too expensive. In addition, researchers need the ability to combine data in a single place temporarily for cleaning, integration, and analysis for projects or community-wide collaborations (e.g., DOE's National Virtual Biotechnology Laboratory research and response to COVID-19).

A unified data framework is not necessarily a single data center but rather a distributed data mesh with a computing capability enabling users to avoid downloading data and instead process datasets where they reside or easily move them to where they can be processed.

4.1.7 Use of New Technologies

The uptake of new technologies such as artificial intelligence (AI) and quantum computing is currently limited. Fast-paced changes occurring within these domains require a significant level of expertise to use these tools correctly and to their full benefit. Key underlying infrastructure barriers limiting greater use of these technologies include (1) identifying suitable dataset collections large enough to train models reliably and (2) providing either the compute capacity to train models on large-scale data *in situ* or the ability to transfer data to a suitable computing platform easily. New AI trends, such as powerful foundation models for science, require even more data ranging from a billion to trillion parameter datasets that need to be vetted for correctness and potential biases.

The research community would benefit greatly from the sharing of models and tools that optimize hyperparameters, enable *in situ* federated learning across different data sources, or assess model fidelity and uncertainty. Support in leveraging quantum technologies on their own or in hybrid classic/quantum approaches would lower the barriers to their use.

4.2 Data Infrastructures: Current State of the Art

4.2.1 European Efforts

Unified data infrastructures are not a new concept. In 2002, the United Kingdom's Natural Environmental Research Council (similar to BER environmental science) began development of its unified Data Grid, which still operates today as NERC Data Catalog (eds.ukri.org/services/find-data). This infrastructure enables data access and metadata search across different archives. Around the same time, the Earth System Grid Federation (ESGF; esgf.llnl.gov) developed as a partnership between BER and Europe. ESGF and Earth Science Information Partners (ESIP), which formed later (www.esipfed.org), are pivotal components in climate and Earth science research, facilitating seamless data management, dissemination, and collaboration. ESGF provides a distributed infrastructure for vast climate model outputs, while ESIP promotes data-sharing best practices and standards for interoperability.

In 2006, the European Strategy Forum on Research Infrastructures called out for the first time the need for treating data as research infrastructure and asked for the creation of a data fabric (ESFRI 2006). Many projects followed suit, leading to the European Open Science Cloud (EOSC; eosc-portal.eu), which aims to provide a unified data infrastructure across many different sciences. In its first implementation phase, EOSC provides a marketplace for scientists to discover data, tools, and resources across Europe. As a precursor, several community-based data fabrics relevant to BER were created, such as C-SCALE for Earth system science (c-scale.eu/fedearthdata/) and ELIXIR for life sciences (elixir-europe.org).

4.2.2 Mission-Specific U.S. Government Efforts

In the United States, the National Science Foundation (NSF) in 2021 began funding efforts to build a general National Scientific Data Fabric (nationalsciencedatafabric.org). These efforts, which are in the development phase, aim to link experiments, computing, tools, and data. The National Institutes of Health (NIH) is pursuing its NIH Cloud Platform Interoperability effort to create a federated genomic data ecosystem (datascience.nih.gov/nih-cloud-platform-interoperability-effort).

There are several other community-specific unified data infrastructure efforts in the United States:

- BER's KBase is an analysis platform that grants users access to data and computing at geographically distributed resources through a Jupyter Notebook interface. Its main data repository, however, is a centralized integration platform.
- BER's National Microbiome Data Collaborative (NMDC) centralizes microbiome data, promoting FAIR principles to accelerate environmental microbiome discoveries.

- BER's Joint Genome Institute (JGI) and Environmental Molecular Sciences Laboratory (EMSL) user facilities provide long-term preservation and access to data and make contributions to national repositories.
- NSF's Long Term Ecological Research Network (LTER) offers a comprehensive data infrastructure capturing diverse ecological data over extended scales, ensuring consistency and accessibility through the LTER Network Data Portal.
- The AmeriFlux network is a collection of long-term, eddy-covariance flux stations that measure ecosystem carbon, water, and energy fluxes across the Americas. The AmeriFlux Management Project works to provide open access to these data in formats that are consistent, standardized, and easy to use. The U.S.-based component of AmeriFlux is part of a broader international consortium of flux researchers who have built an interconnected data infrastructure; the U.S. collaborators lead the way in creating standardized datasets.
- The new National Virtual Biosecurity for Bioenergy Crops Center is a collaborative research platform that brings high-performance computing, high-throughput computing, and AI training and use together with large-scale, temporary, and multiscale data collections and applications.

Different from other unified data infrastructures, data storage, cataloging, and access, all these solutions are centralized rather than distributed.

In 2020, DOE's Advanced Scientific Computing Research program (ASCR) began developing an Integrated Research Infrastructure (IRI) across all DOE facilities and critical capabilities (e.g., experiments, observations, data, and computing). ASCR awarded the centerpiece of its IRI concept—the High Performance Data Facility—in autumn 2023, kickstarting a community consultation phase. Although concepts have been developed, the first IRI testbeds are only emerging and currently restricted to ASCR facilities. A broader inclusion of compute and data resources stewarded by other DOE programs and at other agencies has yet to start.

4.2.3 Commercial and Open-Source Solutions

In addition to mission-specific data infrastructure efforts, recent commercial and open-source tools and frameworks may support the creation of unified infrastructures that enable open and cross-institutional data access and workflows. These particularly include software platforms for AI support of data workflows and federated access to data and computing resources. Frameworks such as Tensorflow Federated and Flower Framework extend popular and robust AI libraries like TensorFlow, PyTorch, and JAX to support federated workflows. Others like APPFLx (appflx.link/), FATE (github.com/FederatedAI/FATE), and OPACUS (github.com/pytorch/opacus) emphasize privacy preservation and secure computing protocols. Commercial (or semi-commercial) products, such as IBM Federated Learning and NVIDIA CLARA, leverage tight integration with a company's tools and platforms.

A common shortcoming of current AI tools and platforms is that they require moderate to extensive familiarity with AI model deployment, command-line usage, software and package management on high-performance computing (HPC) systems, and network and system administration. These barriers to entry may make existing tools unsuitable for noncomputing experts. Moreover, many tools are platform-specific, lack community support, or are partially or fully closed-source, which together may preclude their use in broad, public, and multi-institutional integrated infrastructures. Furthermore, new model types, such as large language or foundation models, require extreme amounts of compute power for their training. These types of resources are not available to or affordable for most researchers.

In summary, there are no current off-the-shelf solutions for unified infrastructures that could address the barriers identified by the workshop community. However, BER could learn from, adopt, or customize many existing solutions and ongoing efforts. For example, EOSC offers a marketplace concept that could address the need for a unified search capability, data standardization, and accessibility. ELIXIR and C-SCALE can offer ideas for more customized services for BER researchers. NSF and ASCR efforts to develop nationwide data fabrics or integrated research infrastructures would be ideal to address needs for cross-facility interactions, unified data and compute capabilities, and changes in existing working practices. Section 4.3 discusses specific steps that can be taken.

4.3 Safe, Secure, and Trustworthy AI in the Context of a BER Unified Data Infrastructure

AI has emerged as a hugely transformative and fast-changing technology for scientific discovery. It has already impacted many areas of BER science through advanced data analytics, fast surrogates, and experimental automation. Rather than examining the many ways in which AI can transform BER scientific discovery, this section will discuss how a unified data infrastructure could assist in leveraging AI for such a transformation and how AI may be transformative for the unified infrastructure itself.

AI models have a significant range of capabilities and can be impactful in many different application areas, from basic pattern recognition in scientific data to the predictive capabilities of models trained on scientific data and numerical models. Optimal results depend on a variety of key factors, including the right choice of model type or combination for a specific task, quality of training data, ability to optimize hyperparameters quickly and correctly, and a reliable and performant execution environment.

AI models are a rapidly evolving field of science. Large language models (LLM), such as those used for applications like ChatGPT, and foundation models have emerged as powerful scientific tools over the past 2 years, driven by the availability of large-scale data and compute power. DOE national laboratories and other research institutions just established a Trillion Parameter Consortium (tpc.dev/) to create a new generation of foundation models with extensive and versatile capabilities. Development and use of these models rely on the same factors as basic AI models but require even larger training-data volumes, exascale compute resources, and sophisticated training frameworks to harness data and compute power.

In October 2023, the White House released an Executive Order on "The Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence" (U.S. White House 2023). The order notes that although AI is incredibly powerful, it also has the potential to cause harm or produce erroneous results if used without safeguards and knowledge. A few possible pitfalls include data bias, uncertainty of results, and "hallucinations" (i.e., highly plausible but incorrect information generated by AI models that is presented as fact). To counteract these potential downsides, U.S. government agencies including DOE are developing strategies, policies, and practical frameworks to ensure the safety, security, and trustworthiness of AI solutions. However, this effort represents a research field that is evolving as quickly as AI itself, and many changes in guidance and tools can be expected over the next decade.

A BER unified data infrastructure could play a pivotal role in putting BER researchers at the forefront of leveraging AI design and use to solve the program's most pressing science challenges. This goal can be achieved by using the infrastructure as a central distribution and resource for policies, best practices, training, and validated tools. In addition to accelerating AI uptake by BER scientists, this approach would also help ensure the integrity of BER research. Moreover, by providing the needed data and compute resources, BER could make these technologies accessible to minorities and underserved communities. Finally, AI could play an important role in solving some of the identified barriers to scientific progress described in Ch. 2: Science Opportunities, p. 7, of this report. Federated data analysis at scale, metadata harmonization, and user guidance are just some of the areas in which AI could have a significant impact.

4.4 Required Research, Development, and Policy

Available software and technologies cannot address all barriers to BER research (see Section 4.2, p. 35). As such, this section highlights key steps that BER can take to support the creation of a highly effective unified data infrastructure. Overall, a collaborative environment based on the European marketplace concept is envisioned in a platform for exchanging services, communication, and data. The subcommittee has identified several priority areas to facilitate the development of such a marketplace. These priorities fall under three critical components: (1) policies that will support a BER unified data infrastructure, (2) development of components based on full or partial existing solutions, and (3) research for areas in which current technologies and solutions cannot overcome the identified barriers.

4.4.1 Policy

Data needs to be standardized and accessible across facilities for easy sharing; this requires metadata preservation and sharing among different communities to account for different naming conventions. Also essential are policies that support collaborative data access, standardization, and integration; enforce inclusive, community-wide governance structures for the unified data infrastructure; and provide researchers with a single sign-on capability across all connected BER facilities and projects. Practical policies in support of safe, secure, and trustworthy AI will help guide the BER research community.

4.4.2 Development

There are three priorities for infrastructure development:

- An effective, scalable search engine to discover suitable datasets, workflows, parameters, tools, frameworks, and algorithms.
- A common mechanism for attributing researchers' contributions (e.g., data, metadata, software, and workflows) to highlight particularly collaborative users.
- A common, central, and distributed data fabric that connects all BER user facilities with a standard user experience and interface, metadata, ontologies, workflows, clean and ready-to-use datasets, tools, and frameworks.

The data fabric requires middleware and services to supply a centralized compute infrastructure with access to datasets (with full context) and the ability to support common workflows, standard ontologies, and metadata. Also needed is a unified infrastructure that links and integrates distributed data and computational facilities. The data fabric will require a scalable private cloud solution (i.e., government cloud) that is more collaborative and requires less initial investment than current commercial options for sharing data, workflows, algorithms, code, and tools. Finally, it will require the ability to store data and metadata from small and large producers not currently supported by an existing BER data facility.

An integrated part of the data fabric should be a hub for tools to enable the development, training, and use of safe, secure, and trustworthy AI solutions.

4.4.3 Research

A unified data infrastructure presents several compelling opportunities for research. A key factor in building the marketplace will be compiling exemplar datasets and pipelines, along with expert commentary, from existing facilities and projects as prototypes for infrastructure best practices. Developing AI in support of data infrastructure creation and maintenance (e.g., data recommendations, connections, and transformations) will also be critical, mainly as data volumes grow and multiscale and multidomain data are integrated. Finally, funding for training and outreach is a clear need for teaching people how to use the marketplace. This investment would enable advanced future BER research.

4.4.4 Training, Support, and Documentation

A unified data infrastructure must be introduced by a strong, integrated program for training, support, and outreach that would lower the entry barrier for users as much as possible and help change current working practices.

To help researchers effectively use the platform, the program should provide comprehensive user training and support resources that include documentation, tutorials, and online forums to address common queries. User feedback and needs should be considered and integrated into efforts to continuously improve user experiences.

Support and outreach efforts should encourage user engagement and foster a community around the unified platform. Developers could organize webinars, workshops, or conferences to promote knowledge sharing and research collaboration and provide opportunities for users to contribute to the platform's development, such as through feature requests or user-driven customization options. Additional training sessions and seminars could help educate users about the platform's capabilities and encourage active participation by showcasing its value to researchers' work. A comprehensive communication and marketing strategy also is needed to promote the unified platform to stakeholders and the scientific community. This strategy would highlight the platform's benefits, features, and success stories.

4.4.5 Role of Existing BER Facilities and Capabilities

Existing BER facilities and capabilities could play an impactful role in building and supporting a new unified BER data infrastructure. Programs such as Facilities Integrating Collaborations for User Science (FICUS) are expanding because of demand from the scientific community. Initially a collaborative effort between JGI and EMSL, FICUS represents a unique opportunity for researchers to combine the power of multiple BER user facilities in one proposed research project. Additional opportunities to leverage current capabilities involve new experiments researchers are developing based on cutting-edge resources across the national laboratories. The data generated at these sites needs to be analyzed and preserved. Ensuring support for the entire data life cycle is critical and will require collaboration across all facilities and projects.

Increased Awareness of Data Resources

Researchers are contending with an overwhelming amount of information and have difficulty finding relevant data and resources; many look to their research teams or knowledgeable experts for help with this challenge. New ways of disseminating information about existing resources and projects are greatly needed. BER may need a data and information hub cited by publications and promoted via existing researchers and social media that is more dynamic than the current BER web presence.

Community Engagement and Training

Each BER resource does not need to engage the same shared community members separately to determine their needs. In fact, operating independent outreach and engagement activities can lead to more confusion and frustration as the same scientists are asked for input in different ways from different projects. BER resources should be encouraged to engage in coordinated outreach activities to help guide the development of infrastructure that is of highest priority to the community. Furthermore, labor-intensive education and training activities should be orchestrated collaboratively, when possible, to achieve impact through consistency.

Metadata Alignment

Some of the metadata challenges identified in this report could be addressed through a common BER sample registration system. Many BER projects have engaged in efforts to apply standards to the metadata associated with scientific data assets, and each facility has systems in place that work well for managing data and metadata for their experiments, but the systems are different. The same is true for how each facility, repository, and project accepts and registers samples. The use of different sample identifiers makes finding all the digital objects associated with the same sample difficult when that data live in different systems. This challenge is particularly acute when combining environmental sample identifiers [e.g., International Generic Sample Number (IGSN)] and biological sample identifiers (e.g., BioSample).

Several BER resources are working together to address this issue. EMSL, JGI, KBase, NMDC, and the Environmental System Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE) are developing a sample harmonization strategy that can bridge the gap between biological and environmental sample metadata. NMDC has a prototype metadata submission system that addresses a key gap in functionality where users can enter metadata and validate it against existing standards in real time. The submission portal has been developed with input from ESS-DIVE, EMSL, JGI, and KBase. This prototype could be expanded to include additional metadata standards for samples, and BER could direct all users to this resource for sample registration. With the sample identification barrier removed, all BER resources could leverage a common sample ID as the method to identify data from a common sample. In addition, users of each resource would have access to high-quality contextual information about the sample, such as where it came from and how it was collected. Without a centralized effort in this area, determining whether two datasets are derived from the same physical sample will remain difficult.

Standardized Distributed Workflow Execution

Robust, resilient, and distributed workflow execution is required to create an abstraction where scientists need not know where their data reside or their computing is executed. Implementing standard workflows at scale across BER and ASCR computing infrastructure is possible, provided scientists create a protocol (e.g., Workflow Description Language, Snakemake file, or Common Workflow Language), a workflow execution environment specification (e.g., Docker container), and use a common resource like Globus for data transfers.

This capability hinges on clear, standardized user authentication and access control policies among computing facilities. JGI, EMSL, NMDC, and ASCR's National Energy Research Scientific Computing Center have prototyped a system for distributed workflow execution across Lawrence Berkeley National Laboratory, the cloud, and Pacific Northwest National Laboratory. This system is used in production by JGI analysis teams and can offer lessons learned regarding policies and technology choices needed for a geographically distributed system, such as the planned ASCR IRI.

BER does not need a single distributed workflow system, but policies that create standard, stable interfaces to computing and data resources are essential (e.g., Superfacility application programming interface) so that developers can focus on enhancing the user experience to improve scientific productivity. These standards would benefit all projects developing server-side computing, including KBase, ESGF, ARM, the MultiSector Dynamics–Living, Intuitive, Value-adding, Environment (MSD-LIVE), and those engaged in large amounts of data processing or analysis.

Data Transfer with Context

Users of BER data systems regularly download or transfer hundreds of thousands of files. Once downloaded, it is incumbent upon the user to track where the data came from and provide appropriate attribution upon publication of a result involving the data's reuse. To facilitate data attribution, KBase staff have developed a standard for transferring appropriate credit or attribution information with data files. KBase and JGI also are collaborating on a demonstration project in which data discovered through JGI are transferred to KBase along with validated provenance. Additional validation for agreed-upon core BER metadata can be added to the system. For example, if BER resources leverage the same sample metadata schema, then the transfer service can also ensure that the sample information is valid and moves with the file. Executing this effort requires (1) policies to align core standardized BER metadata; (2) application programming interfaces from each resource that provide this information; and (3) user engagement to ensure solutions work for the scientific community. Thought must be given to the range of potential cases because tracking information for a single file from a single study is far easier than tracking thousands of files.

Shared Staffing Model

One of the most difficult resources to recruit, develop, and retain is a highly skilled workforce. There is high demand for staff who can design and build high-quality software and hardware infrastructure to support scientific productivity. Software infrastructure built with a user-centered approach reduces the learning curve for researchers and improves accessibility. User-centered software development and design skills are hard to recruit and retain at the national laboratories. Existing facilities should be encouraged to collaborate on common software and hardware infrastructure to gain efficiencies in staffing.

4.5 Five-Year Recommendations

BER has a rich tapestry of data services provided by its user facilities, data repositories, and major projects. With their well-curated experimental and observational data and computing capabilities, these resources form an excellent starting point for a BER unified data infrastructure.

The focus of the first 5 years of creating this infrastructure should be on integrating currently independent BER resources and adding key service and data capabilities that enable an easy and integrated user experience. Furthermore, creating an effective, unified research environment will require a fundamental rethink of the socioeconomic-technical underpinnings of how BER manages data and computing resources, moving away from single facility and focused user communities to a broader, welcoming, and integrated research infrastructure for BER science. To this end, the subcommittee recommends the following nine actions:

1. Develop policies to require the harmonization of user IDs, authentication, and authorization across BER facilities.

2. Establish an inclusive, community-wide governance structure for the developing BER unified data infrastructure.

3. Establish a BER marketplace where BER scientists can discover data, tools, services, and resources. Over time, new capabilities for data sharing, access, transfer, and analysis need to be added, emphasizing those that work across different facilities and communities. Adding data and tool provenance methods, usage statistics, and attribution will be essential. To catalyze broader data use, the unified data infrastructure should include a component that assists researchers in identifying valuable datasets they may not have found in their searches (i.e., a BER recommender engine).

4. Encourage the harmonization of metadata and data formats across domain subject areas, including transformation and integration pathways, starting with existing BER facilities and services. This is a major effort and should initially focus on BER priority research areas, such as the new Biopreparedness Research Virtual Environment (BRaVE) projects, supporting their multidisciplinary science. Community workshops to develop this harmonization will help. These workshops will help to gradually make using the data resources available in the BER unified data infrastructure easier. Developers can leverage machine-learning tools to clean or identify metadata.

5. Develop standardized, sharable, and domain-specific workflows that support easy use of distributed resources and provide excellent starting points for customized, project-specific solutions. Over time, the marketplace should identify and make available for easy use curated

and standardized model-ready datasets and provide examples of well-annotated datasets for others. Qualified and graded datasets will build trust in the BER community and encourage broader usage.

6. Make BER's significant scalable computing capabilities seamlessly accessible and integrate them with the program's data services. Additional resources (e.g., from ASCR facilities) should be included in this seamless integration. Also consider new resources that enable the collection and curation of large datasets of interest to many in the BER research community and combine these with computing capabilities to collaboratively clean, curate, and analyze these data and share outcomes.
7. Build an AI hub as part of this new unified data infrastructure that provides help and practical support for the BER science community, allowing researchers to exchange and collaborate on best-practice development and community standards while learning about the latest developments.
8. Build national and international collaborations to benefit from the expertise and experiences of others, including leveraging existing or planned investments (e.g., ASCR's IRI effort). Within this ecosystem, define a clear role for BER, where it can deliver well-defined added benefits not only to its own research community but the global one as well.
9. Establish a comprehensive training program.

5 Summary and Recommendations

In response to the first charge question—"to review the existing and anticipated capabilities in data management and supporting infrastructures that are relevant to the breadth of BER science"—the subcommittee offers the following findings:

- BER has a sophisticated set of data infrastructure capabilities that support specific programs and can facilitate the integration of data resources across dedicated user communities (see Ch. 1: Introduction, p. 1).
- However, gaps exist in available data services, some datasets collected through BER programs and projects are not easily accessible, and cross-community integration across different data services has limited support (see Ch. 2: Science Opportunities, p. 7)
- Data service interactions with other agencies is limited and cumbersome outside of DOE.

Based on its review, the subcommittee makes the following observations and recommendations for a strategy for next-generation data management and analysis within a unified BER framework.

5.1 Subcommittee Observations

- BER research is increasingly complex, requiring the integration and study of processes across scales and modalities. However, current BER data infrastructure is not ready to support such efforts (see Ch. 2).
- More could be done to provide underserved communities and minorities with easy access to BER capabilities and encourage them to participate in BER research (see Ch. 3: Workforce Development, Inclusion and Accessibility, p. 23).
- New infrastructure strategies could enhance workforce development and, in particular, support early career scientists better (see Ch. 4: Unified Data Infrastructure and Artificial Intelligence, p. 33).
- Many unified data infrastructure efforts are underway worldwide. While none is ready for adoption yet, BER could learn a lot from these efforts, and useful collaborations could be formed (see Ch. 4).
- Creating a unified data infrastructure requires not only technical developments but also the integration of researchers from different communities, allowing them to communicate and interact with ease (see Ch. 3).
- A complete BER unified data infrastructure is not achievable in 5 years (see Section 4. 5, p. 41).

5.2 Subcommittee Strategy Recommendations

- Pursue a project-driven collaboration strategy between infrastructure developers and researchers (adopt a "build it together" approach rather than "build it, and they will come").
- Identify a select number of high-impact science goals that require a unified data infrastructure to empower early adopters, and, ultimately, affect a culture change across the BER research space.
- Explicitly include targeted outreach in early science demonstrators to reach diverse stakeholders and integrate underserved researchers into the initial design phase.
- Leverage existing BER facilities and data services to build an initial tightly integrated unified data infrastructure. Augment this infrastructure with

a dedicated data facility (can be federated) that combines large-scale data and computing to alleviate the need for BER scientists to download data for integration and analysis.

- Establish a BER marketplace where BER scientists can discover and use data, tools, services, and resources across all BER programs, as well as interact with each other and form new collaborations.
- Support targeted outreach and mentoring as data and tools come online to ensure, from the outset, a breadth of users and awareness of tools and data.
- Support the integration of new technologies, such as artificial intelligence, quantum science, and digital twins, through dedicated training, validations, and verification frameworks.
- Support the incubation of a community-based unified data infrastructure through policies to harmonize user IDs, authentication, and authorization across BER facilities and data services.
- Integrate all new infrastructure into the unified data infrastructure and incentivize participation, which likely requires long-term commitment to host data and access.
- Co-develop a buildout plan, based on the requirements of early community adopters, that heavily leverages unified data infrastructures, such as (1) the DOE Advanced Scientific Computing Research program's Integrated Research Infrastructure High Performance Data Facility; (2) the National Science Foundation's National Scientific Data Fabric; and (3) efforts associated with the European Open Science Cloud, including the European Destination Earth project.
- Regularly review and amend the plan to incorporate the evolving requirements and priorities of communities as they work together in the new BER marketplace.
- Selectively support integration and interaction with other agencies' data frameworks important to BER science. Given the effort that such connections require, target only core partners on a project-driven basis in the first 5 years.
- Develop clear metrics of success for all stages and aspects of the unified data infrastructure program.

Appendix A: BER Advisory Committee Members

Chair

Bruce Hungate, *Northern Arizona University*

Members

Caroline Ajo-Franklin, *Rice University*
Cris Argueso, *Colorado State University*
Sarah M. Assmann, *Pennsylvania State University*
Ana P. Barros, *University of Illinois*
Bruno Basso, *Michigan State University*
Julie S. Biteen, *University of Michigan*
Sen Chiao, *Howard University*
Leo Donner, *Princeton University*
Robert F. Fischetti, *Argonne National Laboratory*
Matthew Fields, *Montana State University*
Ann M. Fridlind, *NASA Goddard Institute for Space Studies*
Jorge E. Gonzalez-Cruz, *City College of New York*
Ramon Gonzalez, *University of South Florida*
Randi Johnson, *Formerly U.S. Department of Agriculture*
Kerstin Kleese van Dam, *Brookhaven National Laboratory*
Sonia Kreidenweis, *Colorado State University*
Maureen McCann, *National Renewable Energy Laboratory*
Xiaohong Liu, *Texas A&M University*
Gerald A. Meehl, *National Center for Atmospheric Research*
Gloria K. Muday, *Wake Forest University*
Dev Niyogi, *University of Texas, Austin*
Himadri Pakrasi, *Washington University in St. Louis*
Kristala Prather, *Massachusetts Institute of Technology*
Patrick Reed, *Cornell University*
Gemma Reguera, *Michigan State University*
Jeremy Schmutz, *Hudson Alpha Institute for Biotechnology*
Daniel Segrè, *Boston University*
Karen Seto, *Yale University*
Matthew D. Shupe, *University of Colorado and NOAA*

Appendix B: Request for Information

Agency: Office of Science, Biological and Environmental Research Program, Department of Energy.

Action: Request For Information

Summary: The Biological and Environmental Research (BER) Program, as DOE's coordinating office for research on biological systems, bioenergy, environmental science, and Earth system science, is seeking input on the need and the structure of a unified data framework that links or integrates existing data activities within BER. Information produced in response to this request may be used by the BER Advisory Committee (BERAC) to help inform and recommend to BER a strategy for next-generation data management and analysis within a unified framework.

Dates: Written comments and information are requested on or before October 31, 2023.

Addresses: Interested persons may submit comments by email only. Comments must be sent to BERACRFI@science.doe.gov with the subject line "BER unified data".

For further information, contact: Dr. Tristram O. West, (301) 903–5155, Tristram.west@science.doe.gov.

Supplementary information: A charge was issued from the Director of Office of Science on October 13, 2022, to the BER Advisory Committee (BERAC) to (1) review the existing and anticipated capabilities in data management and supporting infrastructures that are relevant to the breadth of BER science and (2) recommend a strategy for next-generation data management and analysis within a unified framework. The Director's charge letter may be found here: https://science.osti.gov/ber/berac/Reports/Current-BERAC-Charges. Information collected through this request for information, in addition to other informational sources, may be used by BERAC to recommend strategies to further integrate and strengthen BER's data infrastructure in support of BER research. It may also be used by the BERAC in fulfilling its October 13, 2022, charge from the Director of the Office of Science to recommend a strategy for next-generation data management and analysis within a unified framework.

Request For Information

The objective of this request for information is to gather current and future science questions within BER's mission space that would require a more integrated data infrastructure for data access, processing, and use spanning more than one research area. Current BER research areas are provided online: https://science.osti.gov/ber/Research. Supported research includes Atmospheric Science; Earth and Environmental System Modeling; Environmental Science; Bioenergy and Bioproducts; and Plant and Microbial Genomics. Current data archives and activities that support BER research areas include, but are not limited to, ARM https://www.arm.gov/, ESS–DIVE https://ess-dive.lbl.gov/, ESGF https://esgf.llnl.gov/, KBase https://www.kbase.us/, NMDC https://microbiomedata.org/, MSD–LIVE https://msdlive.org/, and JGI https://jgi.doe.gov/.

Information is specifically requested on how a more unified data infrastructure may better facilitate current or future science questions, and what components or technologies are needed to develop a more unified data infrastructure. Answers or information related, but not limited, to the following questions are specifically requested:

1. Do you conduct research in one of the BER research areas (i.e., Atmospheric Science; Earth and Environmental System Modeling; Environmental Science; Bioenergy and Bioproducts; or Plant and Microbial Genomics) and, if so, which area(s)? Please limit additional detail on your area(s) of research interest to a brief paragraph.

2. What new or existing research areas might benefit from improvements in data availability or access across research areas, potentially enabling scientific breakthroughs—and why?
3. What data improvements, including those of accessibility and integration, could facilitate new or existing research or scientific breakthroughs?
    a. Are there current data sets that should be linked or integrated into existing data infrastructure to facilitate existing or new research? If so, which data sets should be so linked or integrated and why?
    b. Are there current barriers to accessing or integrating data from (a) different DOE sources (e.g., ARM, JGI, ESS–DIVE, MSD–LIVE) or from (b) different sources separately maintained by DOE and another Federal agency? If so, what are those barriers and how might they be addressed to allow for improved data access and integration?
    c. What data infrastructure improvements would best support model-experiment feedbacks; facilitate data synthesis and analysis for multi-disciplinary research; and enable application of advanced statistical techniques, including artificial intelligence and machine learning? Please include a brief explanation as to how each identified improvement would support each of these listed tasks.
    d. What current barriers need to be addressed in developing a unified infrastructure to promote greater use by a more diverse community of users, with a focus on improving diversity, equity, and inclusion within data usage and application?

While the questions provided above can help guide thinking on this topic, any input is welcome that may assist BERAC in developing a next-generation data infrastructure in support of BER mission science. The information provided through this request will assist in developing specific strategies that the DOE Office of Science may implement.

Confidential Business Information

Pursuant to 10 CFR 1004.11, any person submitting information that he or she believes to be confidential and exempt by law from public disclosure should submit via email two well-marked copies: one copy of the document marked "confidential" including all the information believed to be confidential, and one copy of the document marked "non-confidential" with the information believed to be confidential deleted. DOE will make its own determination about the confidential status of the information and treat it according to its determination.

Signing Authority

This document of the Department of Energy was signed on April 3, 2023, by Asmeret Asefaw Berhe, Director, Office of Science pursuant to delegated authority from the Secretary of Energy. The document with the original signature and date is maintained by DOE. For administrative purposes only, and in compliance with requirements of the Office of the Federal Register, the undersigned DOE Federal Register Liaison Officer has been authorized to sign and submit the document in electronic format for publication, as an official document of the Department of Energy. This administrative process in no way alters the legal effect of this document upon publication in the Federal Register.

Signed in Washington, DC, on April 12, 2023.

Treena V. Garrett,
Federal Register Liaison Officer
U.S. Department of Energy
[FR Doc. 2023–08029 Filed 4–14–23; 8:45 am]
BILLING CODE 6450–01–P

Appendix C: References

ESFRI. 2006. *European Roadmap for Research Infrastructures.* European Strategy Forum on Research Infrastructures. www.esfri.eu/sites/default/files/esfri_roadmap_2006_en.pdf

Todd-Brown, K. E.O., et al. 2022. "Reviews and Syntheses: The Promise of Big Diverse Soil Data, Moving Current Practices Towards Future Potential," *Biogeosciences* 19(14), 3505–522. DOI: 10.5194/bg-19-3505-2022.

U.S. DOE. 2021. *Biological Systems Science Division Strategic Plan.* U.S. Department of Energy Office of Science. genomicscience.energy.gov/wp-content/uploads/2021/09/BSSD_Strategic_Plan_2021.pdf

U.S. White House. 2023. *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.* U.S. White House. www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

Appendix D: Acronyms and Abbreviations

AI	artificial intelligence
API	application programming interface
ARM	Atmospheric Radiation Measurement user facility
ASCR	DOE Advanced Scientific Computing Research program
BER	DOE Biological and Environmental Research program
BERAC	Biological and Environmental Research Advisory Committee
BES	DOE Basic Energy Sciences program
BRaVE	Biopreparedness Research Virtual Environment
BRC	Bioenergy Research Center
BSSD	Biological Systems Science Division
CEDS	Community Emissions Data System
CF	Climate and Forecast
CMIP	Coupled Model Intercomparison Project
CO_2	carbon dioxide
COMPASS	Coastal Observations, Mechanisms, and Predictions Across Systems and Scales
COVID-19	coronavirus disease 2019
DEI	diversity, equity, and inclusion
DOE	U.S. Department of Energy
DOI	digital object identifier
E3SM	Energy Exascale Earth System Model
EESSD	Earth and Environmental Systems Sciences Division
EMSL	Environmental Molecular Sciences Laboratory
EOSC	European Open Science Cloud
EPA	Environmental Protection Agency
ESGF	Earth System Grid Federation
ESIP	Earth Science Information Partners
ESS-DIVE	Environmental System Science Data Infrastructure for a Virtual Ecosystem
EUMETSAT	European Operational Satellite Agency
FAIR	findable, accessible, interoperable, and reusable
FICUS	Facilities Integrating Collaborations for User Science
GCAM	Global Change Analysis Model
HBCUs	historically black colleges and universities
HPC	high-performance computing
iCoM	integrated coastal modeling
IGSN	international generic sample number
IRI	integrated research infrastructure
JGI	DOE Joint Genome Institute
KBase	DOE Systems Biology Knowledgebase
LLM	large language model
LTER	Long-Term Ecological Research Network
ML	machine learning
MSD–LIVE	MultiSector Dynamics–Living Intuitive Value-Adding Environment
MSI	minority-serving institution
NAIRR	National Artificial Intelligence Research Resource
NASA	National Aeronautics and Space Administration
NEON	National Ecological Observatory Network
NGEE	Next-Generation Ecosystem Experiments
NIH	National Institutes of Health
NMDC	National Microbiome Data Collaborative
NOAA	National Oceanic and Atmospheric Administration
NSF	National Science Foundation
RENEW	Reaching a New Energy Sciences Workforce
RFI	Request For Information
STEM	science, technology, engineering, mathematics
UIFL	Urban Integrated Field Laboratory
USGS	U.S. Geological Survey
USFS	U.S. Forest Service