DOE/SC-0180

WORKING GROUP ON VIRTUAL DATA INTEGRATION

A REPORT FROM THE AUGUST 13–14, 2015, WORKSHOP | BETHESDA, MD

U.S. DEPARTMENT OF ENERGY | Office of Science

Office of Biological and Environmental Research

Working Group on Virtual Data Integration

A Report from the August 13–14, 2015, Workshop

Bethesda, Maryland

Convened by

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

BER Program Manager

Justin Hnilo
justin.hnilo@science.doe.gov

Workshop Attendees and Report Contributors

Deb Agarwal, David C. Bader, Thomas A. Boden, Scott M. Collis, Jennifer Comstock, Eli Dart, Paul J. Durack, Ian Foster, Forrest M. Hoffman, Robert Jacob, Philip J. Rasch, Timothy Scheibe, Mallikarjun Shankar, David Skinner, Peter Thornton, Margaret S. Torn, Andrew Vogelmann, Michael F. Wehner, and Shaocheng Xie

Workshop and Report Organizers

Dean N. Williams
Lawrence Livermore
National Laboratory
williams13@llnl.gov, 925.455.4774

Giriprakash Palanisamy
Oak Ridge
National Laboratory
palanisamyg@ornl.gov, 865.241.5926

Kerstin Kleese van Dam
Pacific Northwest
National Laboratory
Brookhaven National Laboratory
kleese@bnl.gov, 631.344.6019

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Suggested citation for this report: U.S. DOE. 2016. *Working Group on Virtual Data Integration: A Report from the August 13–14, 2015, Workshop*. DOE/SC-0180. U.S. Department of Energy Office of Science. DOI:10.2172/1227017.

DOE/SC–0180

Working Group on Virtual Data Integration

A Report from the August 13–14, 2015, Workshop

Bethesda, Maryland

Requirements to Achieve BER's Vision of a Virtual Laboratory: A Next-Generation Data Infrastructure for Climate Science

Convened by

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

Publication Date: February 2016

U.S. DEPARTMENT OF ENERGY | Office of Science

Contents

Executive Summary v
1. Background and Introduction 1
2. Scientific Challenges and Motivating Use Cases 3
2.1 Use Case 1: Collaborative Scientific Discovery Across Discipline Boundaries 4
2.2 Use Case 2: Multisource Observational Data Integration 6
2.3 Use Case 3: Climate Modeling and Model Analysis 7
3. Survey Results 9
4. Data Services Needed to Support Science Requirements 13
5. Advanced Computational Environments and Data Analytics 15
6. Data Centers and Interoperable Services 19
6.1 Earth System Grid Federation 19
6.2 ARM Climate Research Facility Data Center 21
6.3 Carbon Dioxide Information Analysis Center 22
6.4 Other Interoperable Services 23
6.5 Recommended Interoperable Services 23
7. Inventory of Existing CESD Computer Resources, Data Tools, and Services 25
8. Data Services and Monitoring 29
9. Synergies with Peta- and Exascale Computing Hardware 31
10. Network Services 33
11. Participation with Broad Multiagency Data Initiatives 35
11.1 NASA 35
11.2 NOAA 36
11.3 NSF 37
11.4 Opportunities for Coordination 37
12. References 39
Appendix 1. Workshop Findings 41
Appendix 2. Workshop Example Questions 44
Appendix 3. Survey Questions — Overall Ranking 45
Appendix 4. Workshop Agenda 47
Appendix 5. Workshop Participants 50
Appendix 6. Acronyms, Abbreviations, and Terms 51

Executive Summary

This report is the outcome of a workshop commissioned by the U.S. Department of Energy's (DOE) Climate and Environmental Sciences Division (CESD) to examine current and future data infrastructure requirements foundational for achieving CESD scientific mission goals in advancing a robust, predictive understanding of Earth's climate and environmental systems. Over the past several years, data volumes in CESD disciplines have risen sharply to unprecedented levels (tens of petabytes). Moreover, the complexity and diversity of this research data—including simulations, observations, and reanalysis—have grown significantly, posing new challenges for data capture, storage, verification, analysis, and integration. With the trends of increased data volume (in the hundreds of petabytes), more complex analysis processes, and growing cross-disciplinary collaborations, it is timely to investigate whether the CESD community has the computational and data support needed to fully realize the scientific potential of its data collections. In recognition of the challenges, a partnership is forming across CESD and among national and international agencies to examine the viability of creating an integrated, collaborative data infrastructure: a *Virtual Laboratory*. The overarching goal of this report is to identify the community's key data technology *requirements* and high-priority *development needs* for sustaining and growing its scientific discovery potential. The report also aims to map these requirements to *existing solutions* and to identify *gaps* in current services, tools, and infrastructure that will need to be addressed in the short, medium, and long term to advance scientific progress.

Prior to the workshop, a survey was circulated to attendees and their associates. Responses emphasized, in particular, a concern about sustained supply of sufficient computational and storage resources. More broadly, the researchers indicated a need for cross-cutting, integrating solutions that address the full spectrum of data lifecycle issues—collection, management, annotation, analysis, sharing, visualization, workflows, and provenance. The top 10 most-cited requirements were (1) an easy way to publish and archive data, (2) comparison of heterogeneous data types, (3) user support and documentation, (4) access to observational and experimental resources, (5) scientific and computational reproducibility, (6) data movement from archives to supercomputers, (7) unification of single user accounts across DOE resources and facilities, (8) resource reliability and resiliency, (9) intuitive human-computer interaction, and (10) quality control algorithms for data. Responses also indicate that methodologies for knowledge gathering, management, and sharing represent an overall area requiring more community attention.

In addition to the survey, this report recognizes community infrastructure investments that support and enable analysis of massive, distributed scientific data collections and that leverage distributed architectures and compute environments designed for specific needs. The report captures this trend by first describing the scientific challenges in the form of diverse and disparate use cases. These scientific use cases demonstrate and emphasize the need for data services, data centers, interoperable services, advanced computational environments, data analytics, data monitoring, multiagency collaboration, and the evaluation of existing tools and services for potential reuse. Workshop participants discussed existing community infrastructures that can help build CESD's Virtual Laboratory, including (1) the Earth System Grid Federation (ESGF), which primarily serves simulation data to the global climate research community; (2) the Atmospheric Radiation Measurement (ARM) Climate Research Facility's Data Center, which collects and provides observational data on aerosols, clouds, and their impacts on Earth's radiation budget; and (3) the Carbon Dioxide Information Analysis Center (CDIAC), which contains observations of ecosystem-level exchanges of carbon dioxide, water, energy, and momentum at different timescales for sites in the Americas. Although these infrastructures may cover some requirements of the scientific use cases, they are not general enough to address all of them and thus will require enhancements to fulfill CESD's scientific vision.

Along with these use cases and infrastructure descriptions, this report highlights a number of core findings by workshop participants:

- Knowledge capture, management, and sharing are key development areas.

- Additional enabling data capabilities are needed throughout the full research lifecycle—from data discovery, to the handling and treatment of large volumes of multisource data, to flexible tools for data analysis, to scientific and computational reproducibility, to data publication and attribution.
- Achieving community scientific goals requires additional storage and computing resources, along with a common virtual computational environment that conforms to established standards across DOE Office of Advanced Scientific Computing Research (ASCR) computing facilities.
- Identifying, applying, and following key interoperability enablers are all critically important when developing tools for CESD programs and projects. Such enablers include metadata conventions and standards, workflow and provenance capture, and data and visualization protocols.
- An inventory is needed outlining the available data, compute tools, and resources currently used by CESD and its associated research communities. Evaluation and assessment of these shared data, tools, and resources would ease their route to adoption into the integrated data ecosystem.
- A new class of monitoring services for the next generation of complex workflows would be valuable, particularly services that capture metrics on data and software downloads, users, and publications resulting from the reuse of a researcher's data and software by others.
- ASCR computing facilities need a policy for retaining data sets with a useful lifespan that extends beyond supported compute facility programs [e.g., the Innovative and Novel Computational Impact on Theory and Experiment (INCITE) program]. Establishing a single sign-on for authentication and federated access also would ease researchers' use of multiple ASCR computing hardware and resources.
- Advances in current high-speed reliable data movement are necessary for sufficiently meeting CESD data resiliency and backup needs.
- Strengthening partnerships with other national and international agencies is necessary for research community success.
- Data management, stewardship, and curation are ongoing and long-lived functions requiring a strategy that is resilient to continuing hardware and software evolution.

Additional workshop findings are further elaborated in Appendix 1, p. 41.

1. Background and Introduction

The Climate and Environmental Sciences Division (CESD) within the U.S. Department of Energy's (DOE) Office of Biological and Environmental Research (BER) focuses on advancing a robust, predictive understanding of Earth's climate and environmental systems by exploiting unique modeling, observational, data, and infrastructure assets developed and managed by BER. CESD's programmatic interest in obtaining this systems-level understanding is driving the need to integrate data and modeling efforts from multiple disciplines. In 2013, the BER Advisory Committee (BERAC) issued a report titled *BER Virtual Laboratory: Innovative Framework for Biological and Environmental Grand Challenges*, which outlined a high-level concept for a potential effort to address the need to integrate data and modeling activities (BERAC 2013). The purpose of that report was to assist in the development of a clear vision and potential plans for a federated BER Virtual Laboratory. Such a system first would be a unified data construct (e.g., a data infrastructure) where any data could reside and be discoverable and, second, offer a compute environment allowing for rapid prototyping, integration, and validation of model modules.

Emphasizing data infrastructure needs in pursuit of exploiting unique models and observations, CESD released its own strategic roadmap for Earth system science data integration in 2014 (Williams et al. 2014). The report introduced a data ecosystem that integrates into a seamless and unified environment all existing and future distributed CESD data holdings. The roadmap describes a highly coordinated set of data-oriented research activities, with a goal of providing the CESD scientific community with easy and efficient access to all data archives necessary for studying increasingly complex scientific challenges. In addition, the report highlights supporting activities involving (1) metadata compatibility from disparate research projects; (2) fusion of data derived from laboratory studies, field observatories, and model-generated output; (3) server-side analysis; and (4) efficient data storage, pattern discovery, and use of computing facilities and networks supported by the DOE Office of Advanced Scientific Computing Research.

CESD currently supports a variety of simulation and observational data archives, including:

- The data center of the Atmospheric Radiation Measurement (ARM) Climate Research Facility, a DOE national user facility (www.arm.gov).
- Carbon Dioxide Information Analysis Center (CDIAC, cdiac.esd.ornl.gov).
- AmeriFlux Network (ameriflux.ornl.gov).
- Next-Generation Ecosystem Experiments (NGEE)–Arctic (ngee-arctic.ornl.gov), NGEE–Tropics (esd.lbl.gov/ngee-tropics), and various data archives within BER's Terrestrial Ecosystem Science (TES) program.
- Earth System Grid Federation (ESGF), the largest model-derived, ensemble-run data archive used by the international climate research community (esgf.llnl.gov, Williams et al. 2015).

In addition to ESGF and the observation-only archives, a number of test beds associated with observational and other laboratories independently manage model-derived data products, such as those generated by the Accelerated Climate Modeling for Energy (ACME) project (U.S. DOE 2014).

CESD data archives and test beds generally evolved independently of each other to support their corresponding user communities. These archives used domain-specific metadata and data standards for processing, archiving, and distributing their data, and historically there was little need to focus on metadata compatibility and broader connectivity among systems and communities. However, current high-priority BER research questions involve complex data from multiple sources (e.g., physical and biogeochemical interactions). These scientific priorities are changing the status quo because they require closer collaboration among scientists from different disciplines and, in turn, better integration of data, tools, and services from CESD and partner data centers, facilities, and resources (ASCAC 2013).

To assist in the development of a better-integrated environment, CESD conducted a "Working Group on Virtual Data Integration" workshop to lay the groundwork

for a federated BER Virtual Laboratory and CESD's data infrastructure, as described in the BERAC 2013 and Williams et al. 2014 reports. For this workshop and report, key CESD personnel—including project leaders, data providers, lead developers, and many others—came together to discuss key cross-cutting requirements. The hope is that this multidisciplinary effort will forge a robust vision for the future in terms of requirements, solutions, and a prioritized approach to creating needed capabilities.

Questions addressed at the workshop and in this report include scientific gaps and challenges to be addressed in the planning and development phases of the Virtual Laboratory, with an emphasis on data infrastructure and the compute environment. Example questions can be found in Appendix 2, p. 44. This report also establishes key community needs and the required deliverables to address them based on clearly articulated use cases. These use cases were developed by current principal investigators of projects within CESD programs and facilities, including the DOE Environmental Molecular Sciences Laboratory; ARM Climate Research Facility; and the Subsurface Biogeochemical Research, TES, Regional and Global Climate Modeling, Earth System Modeling, Atmospheric System Research, and Integrated Assessment programs.

2. Scientific Challenges and Motivating Use Cases

Any realization of a data and informatics system that serves the needs of BER and its Virtual Laboratory concept must be structured to meet requirements imposed by the science and research activities carried out in support of BER's mission. The particular focus for this workshop was on the data system requirements emerging from multidisciplinary research to understand and predict Earth's climate, internal variability, and response to forcing from human activity. The research community broadly recognizes that modeling, observation, and experimentation are all required to advance a predictive understanding of the Earth system and climate change and that many disciplines and scales of study must be engaged to increase the quantitative knowledge of the Earth as a coupled system.

Fig. 1, this page, shows one view of an iterative process of scientific inquiry, hypothesis formation, real-world observation and experimentation, and modeling designed to increase quantitative knowledge and advance efforts to predictively understand Earth's climate and environmental systems. One interpretation of this process is that it encapsulates a large body of scientific and research effort in terms of what scientists do (teal boxes in Fig. 1) and what they produce (blue boxes). Earth system science is characterized by complexity, in that multiple sets of disciplinary knowledge must be integrated and their interactions grasped across a wide spectrum of spatial and temporal scales, from cellular to planetary and from fractions of a second to millennia. Data and metadata (descriptive information about the data) are core components at each step in this research, given the inescapable diversity in the types of data gathered and the ways in which data are generated, processed, and applied in pursuit of increased predictive understanding.

The workshop did not attempt an exhaustive assessment of the requirements placed on a data and informatics system by the entire scope of science and research activities relevant to BER's climate mission. Instead, participants used a small number of use cases to help identify the most significant needs in terms of a system to support what the scientific community does and what it produces. The use cases presented here are examples. They have enough detail to motivate specific, actionable requirements for a data and informatics system but are not intended to cover the entire programmatic scope or range of capabilities that might be demanded of an operational system. These use cases were designed to represent complex requirements emerging from multi-institutional, multidisciplinary, and multiscale investigations, in the hope that cases of this sort would

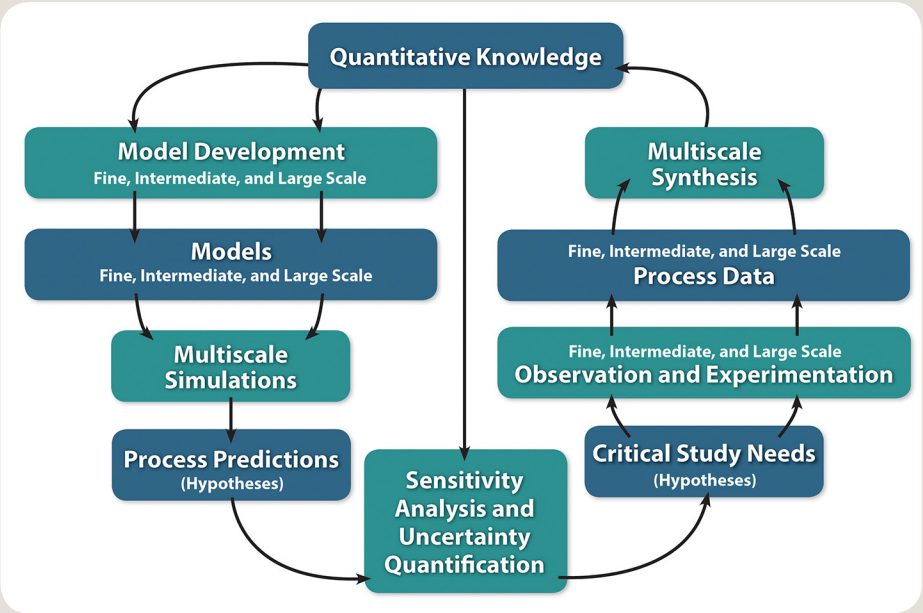

Fig. 1. An iterative approach designed to increase quantitative and predictive knowledge of the climate system through coordinated observation, modeling, experimentation, and uncertainty quantification.

help identify the broadest outlines of a BER-centric climate data and informatics system and its capabilities.

The use cases themselves are shown in plain text, and specific requirements emerging from individual aspects of the use cases are shown as indented italic text.

2.1 Use Case 1: Collaborative Scientific Discovery Across Discipline Boundaries

This use case illustrates the science requirements placed on a data, informatics, and computation system by a collaborative project involving modelers, computational scientists, data scientists, observationalists, and experimentalists.

A synthesis of previously published observational, experimental, and modeling results has indicated that strong interactions among temperature, humidity, and soil moisture control the composition of vegetation communities and the fluxes of carbon dioxide (CO_2) and methane (CH_4) from wetlands in a variety of geographic settings and climate zones.

> *Synthesis studies require indexed access to comprehensive literature and data set resources, with the ability to search and filter by time, geographic location, process-based keywords, investigator, research program, and other fields. Mapping among multiple ontologies or dictionaries is needed to span existing resources.*

A manipulative field experiment is initiated to further explore the influence of long-term warming on CO_2 and CH_4 fluxes in a boreal wetland setting. Pretreatment observational campaigns have characterized the structure and function of the target wetland.

> *The detailed experimental design should be documented in a searchable format so that other researchers can find the intended measurements, manipulations, and other background information even before the experiment is running. This documentation should include points of contact for each element of the experimental design so that questions about potential collaborations can be effectively directed and addressed.*

> *Pretreatment observations and characterizations are critical for modeling studies and should be planned and catalogued with as much forethought and attention to detail as the eventual experimental results. Iteration with modeling groups and other experimental efforts is essential to ensure comprehensive pretreatment observations, since opportunities are necessarily limited once treatments are under way.*

Site-level modeling in advance of a field manipulation indicates that imposed warming will interact strongly with over-winter snowpack, generating a seasonal pattern of positive soil temperature anomalies that are strongest in summer and weakest in winter.

> A priori *modeling is as crucial as pretreatment measurements to the eventual success and applicability of the collaboration. The bootstrapping nature of this kind of work in a newly developed location means that the data system will need to support a range of synthesis efforts to gather existing driving data, interpolation, and gap-filling methods to make data as relevant to the experimental location as possible. Frequent iteration with the experimental team is necessary to ensure that simulations reflect experimental plans. Simulation results need to be made available in searchable formats so that the experimental team can query potential outcomes and ideally be able to establish what-if scenarios to assess details of the experimental design and measurement plan.*

Additional modeling across a latitudinal gradient of wetland sites indicates a complex relationship among warming, seasonal patterns of soil temperature and moisture, and net changes in greenhouse gas budgets.

> *In addition to intensive modeling at the experimental site, extensive multisite modeling at similar locations, potentially including other observational or experimental sites, is a critical step in understanding the relevance of site-level findings. A multisite simulation capability becomes essential in organizing inputs and outputs as the number of additional sites increases from a few to tens or more. A data and informatics system capable of organizing inputs and outputs and allowing evaluation against a range of observational data types could bring great efficiency to this process.*

As detailed experimental results emerge from the long-term warming experiment, short-term sampling is being carried out across a latitudinal gradient and

both experimental and observational data are being deployed in an uncertainty quantification (UQ) framework to evaluate model predictions.

A system that monitors and reports experimental findings in near-real time can enhance the ability to collaborate with an interdisciplinary and multi-institutional team. Continuity with the indexing functions for pretreatment data and the ability to issue problem reports and updates and browse analytics for quick views would all be useful features of such a system.

Field campaign–style sampling across a range of coordinated sites will be a common science requirement in collaborative projects. The ability to replicate a data model with applicability to many sites improves the efficiency of later synthesis and analysis.

Comparison of model results to observations or experimental data imposes a broad set of science requirements. Necessary aspects of a model-data evaluation system include handling of missing values, unit conversions, spatial and temporal aggregation, scaling of measurement uncertainties in space and time, relative weight assignment to multiple independent observations of the same quantity, and definition of model skill metrics. The high dimensionality of model and observational data sets challenges traditional analysis tools and methods, and advance analytics with responsive user interfaces would accelerate new knowledge generation in this area.

At the same time, an effort is under way to integrate representation of wetland thermal hydrology; soil biogeochemistry; and vegetation structure, function, and dynamics into the land component of a coupled Earth system model (ESM). Synthesis studies and multisite modeling suggest that improved process representation could allow the global model to capture the hypothesized mechanistic controls on wetland carbon cycle and surface energy budgets.

New model development requires a design process that, ideally, is as comprehensive and deliberate as the design of new observational or experimental efforts. A broadly capable data, analytics, and computational system would enable this design process through synthesis and evaluation tools. It also would capture the results of the design process in design documents that describe new process representations, model inputs and outputs, and science requirements for parameterization and evaluation of data and analytical resources.

A set of new parameters for the global model must be estimated on the basis of previous literature estimates, new cross-gradient observations, and extensive data collection at the experimental site. A Bayesian UQ framework is deployed to assess model sensitivity to parametric uncertainty, and the most critical parameters are estimated based on multiple independent observational and experimental constraints, each accompanied by uncertainty estimates.

Formal parameter estimation places significant demands on the computational system because a large number of carefully regulated simulations are typically required. A system that cross-references uncertainty estimates on observational and modeling results is also needed to ensure that empirical constraints are applied appropriately. Analysis of large and multidimensional model outputs is required to interpret UQ results. Filtering of sensitivity analysis results produces a reduced set of parameters for formal estimation, but these results can vary in space and time, placing high demands on the analysis framework and requiring engagement of expert knowledge.

The new global model is first evaluated at site and regional scales against withheld observations and then exercised at the global scale. Global-scale simulations include a series of offline simulations driven with observed surface weather, followed by fully coupled ESM simulations covering historical and future periods, out to year 2100, under a variety of socioeconomic forcing assumptions.

Strict model evaluation using withheld data and cross-validation methods provides a conservative estimate of model performance and should be enabled in addition to the more sophisticated Bayesian estimation methods. Challenges here include a diversity of spatial and temporal scales in the available observational and experimental data and the need to aggregate observations or disaggregate model results to make meaningful comparisons.

This is an area of the collaborative use case where existing technologies and practices are already quite mature. Globus, ESGF, and the Coupled Model Intercomparison Project (CMIP) archive, to name a few, have resulted from long-term investment in this area. Reproducibility of results and model configurations in the context of large assemblages of sophisticated simulations are also important aspects of the global-scale simulation problem, but so far they have received less attention and have fewer existing solutions.

Single-factor and multifactor simulations are performed to evaluate the influence of the new wetland model and parameterizations on global-scale climate-biogeochemistry feedbacks.

Complex evaluation methods are employed to assess system feedbacks and signal detection and attribution. Science output and knowledge growth would be enhanced if these complex workflows were incorporated into a broadly capable system.

Results from the global simulations are periodically evaluated against new findings from the experimental site as long-term effects emerge under the warming manipulation.

The overall science objective of hypothesis testing needs to be accommodated in an integrated system. Evaluating global modeling results against newly emerging experimental and observational results is a crucial step in that process. A capable system could help in the synthesis of these periodic evaluations, leading ultimately to refined hypotheses and new process investigations.

2.2 Use Case 2: Multisource Observational Data Integration

This use case highlights the complex relationships that exist among researchers with respect to data collection, stewardship, ownership, and distribution. It also highlights the need for any data and informatics system to maintain transparent records on data provenance and clear guidance on attribution of credit for various stages in the data and project lifecycle.

Dr. Sally Fields is working at the Harvard Forest and is beginning a nitrogen addition experiment. In laying out this experiment, she has decided what she is going to measure and what metadata she wants to record. She does a quick search for standard templates and metadata specifications and does not find any existing standards for the type of experiment she is doing.

Metadata standards and metadata searching capability with interoperability among multiple data centers within various institutions, agencies, and nations are core capabilities that enable all use cases.

Dr. Fields begins the experiment and performs quality assurance/quality control (QA/QC) daily as part of her monitoring effort. When she completes the experiment, she does not have time to analyze the data further or write a paper using the experimental results because she has to teach a class. The data are currently stored in an Excel spreadsheet. She created the QA/QC flags she needed to indicate the various situations as they occurred and built a key for the flags as she went along. She also took soil cores, which were analyzed by a commercial laboratory until it went out of business and she had to switch to another lab. She has received the data from both labs. Although the two labs used different methods to analyze the cores, both generated total nitrogen content values for the cores, although one was by weight and the other by volume.

Any metadata standards need to be as comprehensive as possible but also flexible to accommodate new situations and data types. QC information is a necessary component of data and metadata, especially as data products are shared in larger communities where first-hand knowledge of limitations is the exception rather than the rule.

Dr. Fields attends the North American Carbon Program conference and meets John Flux, who measures similar data in Canada; Dr. Marsh, who has LiDAR data for Harvard Forest; Dr. Nitrogen, who measures leaf-level nitrogen at both sites; and Dr. Cycle, who specializes in modeling nitrogen. They would like to work together to do a model validation using the data from the two sites. Dr. Marsh is part of a large data repository and analysis center, and he offers to host all the data they use for the validation at his site. Dr. Fields is a bit worried about this, since she does not want the

data to be available outside this collaboration until she writes a paper about her results for her tenure case.

> *Productive collaborations require recognition of diversity in requirements and expectations for data sharing, and a broadly capable system needs to both record and protect the interests of multiple parties.*

After they get started, Dr. Flux decides that he does not have the time to contribute to the writing of the paper but that the group is still welcome to use his data as long as he receives credit for it; he has not yet written a paper based on the data and is concerned about continued funding for data collection. (More generally, he is making his data available to any interested user via Dr. Marsh's system but would like credit for his contribution.) He is also concerned that there might be QA/QC problems with the data and would strongly prefer to see any results based on the data before they are published. Dr. Nitrogen has already provided his data set to AmeriFlux, and it is available there with a digital object identifier (DOI), so he asks the team to use that version and cite the DOI. Dr. Cycle has not yet published a paper about her model and is not yet ready to release it, so she would prefer to run all the validations on her own cluster.

> *The lifecycle of data and information in a multi-partner collaboration can be complicated, and provenance information that shows previous, current, and planned future stages in the life-cycle for a given data set needs to be maintained and amended as the collaboration proceeds.*

2.3 Use Case 3: Climate Modeling and Model Analysis

This pair of use cases deals specifically with generation and analysis of large volumes of data from single or multiple ESMs running one or more simulation experiments, sometimes with multiple ensemble members per experiment. Different types of analyses invoke different data storage, handling, and processing requirements.

2.3A Model Intercomparison for Study of Extra-Tropical Cyclones

Dr. Bigdata is conducting a study of how the frequency and severity of extra-tropical cyclones will change under different future carbon scenarios. The source data set is the CMIP, phase 5 (CMIP5) model data. The timescales necessary for tracking evolving weather systems are relatively short, so 6-hour data sampling is required, resulting in an initial data set much larger than those assembled by most scientists (many tens of terabytes). Only a small subset of the data is actually needed for the analysis, but none of the globally distributed data centers that host the CMIP5 archive have an existing analysis tool for this purpose.

> *Although the existing network of climate model data centers supports many routine search and subsetting capabilities, new and innovative analyses are constantly emerging. These analyses can place unforeseen demands on the existing data systems, meaning that a flexible and configurable capability is needed in addition to the standard hosting facilities. This capability may take the form of a prototyping environment, but the storage and processing requirements for new prototype analyses can be very large.*

Dr. Bigdata identifies the model variables required for the analysis and submits a query to the ESGF data infrastructure. This request results in a set of Globus data transfer jobs that deliver the data from the ESGF data infrastructure to a file system at a DOE Office of Advanced Scientific Computing Research (ASCR) computing center. Once the data arrives on the file system, Dr. Bigdata then runs a secondary processing code that requires the massively parallel environment of a national computing facility. The result of this secondary code is a high-value data set of substantially reduced scale, which provides significant leverage for all downstream analyses, especially if other scientists can publish the derived data set with metadata that facilitates interpretation of the data.

> *The prototyping and secondary analysis environment needs to be close to high-throughput data transfer networks and needs access to high-end computational power. The ESGF data infrastructure must also be able to support the necessary large-scale data transfers to a computing facility, which has sufficient capability to run the analysis. New value-added data products need to enter a data lifecycle tracking system that ensures proper metadata, indexing, and attribution are generated and retained.*

Dr. Bigdata's colleague, Professor Sandy Katrina, is studying the effects of climate change on tropical cyclones. Instead of obtaining data from a distributed CMIP5 data set, she creates a modestly large set of multidecadal simulations by running the ACME climate model at a 25 km horizontal resolution. Upon completion of her simulations, she runs the same secondary processing code to identify storms and their tracks. She then uses the track data to query a large three-dimensional subdaily data set to examine changes in storm structures.

> *Many climate modeling applications require dedicated analysis and data-reduction capability at the high-performance compute sites where models are run. In addition to compute capacity for data reduction and batch-mode analysis, interactive visualization capabilities can accelerate the identification and extraction of new knowledge from large multivariate data sets.*

2.3B Three-Dimensional Ocean Analysis

Dr. Lotte Malte-Modele is studying the global ocean and how oceanic variability and change might evolve in response to a series of future CO_2 scenarios. The source data set is from the CMIP5 and CMIP, phase 6 (CMIP6) models. Because of local storage limitations, Dr. Malte-Modele would like to undertake a considerable data reduction on the ESGF nodes where the data reside, thereby reducing the total local footprint required to store the analyzed outputs and decreasing data transfer volumes. As part of the data reduction, Dr. Malte-Modele needs to analyze data on the native grids provided by the modeling centers, often performing calculations that require careful treatment of computed transport. For this, she needs specialized software that is "grid aware" and considers cell volume weights during calculations.

> *To ensure scientific validity and publishable results, analysis software must meet exacting technical requirements. General-purpose software may not be fit for special-purpose analyses, and a data and informatics system needs to be explicit about the capabilities and limitations of default software while accommodating special-purpose software.*

Dr. Malte-Modele identifies the ocean variables required for the analysis and submits a query to the ESGF archive to obtain a list of all available data located across the federated archive. She then constructs an analysis script using the Ultrascale Visualization–Climate Data Analysis Tools (UV-CDAT) analysis package, which is co-located with the data on each ESGF node. Thanks to local resources available on ESGF nodes, this task is completed within a couple of hours. These reduced data are then transferred to local storage using a series of Globus data transfer jobs initialized at one of the ASCR computing centers. Using local software stacks, Dr. Malte-Modele then undertakes the final stages of her research using an array of analysis and graphics tools to prepare publication-ready figures.

> *Some challenging data analysis and handling requirements are already being met by existing systems; therefore, a BER-centric effort for data and analytics need not start from scratch.*

3. Survey Results

To prepare for the workshop, the organizing committee conducted an online survey of researchers supported by CESD. The survey aimed to ascertain what scientists believed were the greatest needs for additional support. The request for feedback was sent not only to workshop participants, but also to CESD principal investigators. Most questions asked researchers to indicate on a scale from one to six if they saw a need for additional support (one indicated no or little interest, while six indicated a highly important area for development). The survey calculated average values for each question across all responses (i.e., a value of 4.79 for a particular topic would indicate that most participants would rate the topic as high or very high interest). It also calculated the percentage of participants that gave a topic a particular rank (e.g., 41% ranked this as very high).

The survey asked the respondents to identify themselves as data providers, resource providers, software developers, climate modelers, or data analysts (see Table 1, this page).

More than 40% of responding scientists saw access to sufficient computational and storage resources as a very high need (see Table 2, p. 10). Also notable was their emphasis on more reliability and resiliency in the resources and services provided to them (34% identified this as their highest need) and access to sufficient observational and experimental capabilities (26%). Data and software resources were identified as the most difficult to discover, with 40% of respondents stating that they might need hours or days to find what they needed. Related to this challenge were requests for more user support for data

> Knowledge capture, management, and sharing are key development areas.
>
> Additional enabling data capabilities are needed throughout the full research lifecycle—from data discovery, to the handling and treatment of large volumes of multisource data, to flexible tools for data analysis, to scientific and computational reproducibility, to data publication and attribution.

Table 1. Self-Identification Categories for the 75 Scientists Who Responded to the Survey Request

Scientific Background	Description	Total
Data provider	Provides data and metadata (describing the data) to the community. Also responsible for data quality. Associated with climate modeling groups and data centers.	32
Resource provider	Provides hardware and software resources at high-performance computing facilities.	4
Software developer	Develops stand-alone software for the climate community. Also known as a computer programmer, application developer, and system software developer.	6
Climate modeler	Develops quantitative methods to simulate the interactions among the important drivers of Earth's climate, such as atmosphere, oceans, land, and sea ice.	15
Climate model data analyst	Analyzes output to understand simulation and observational output for knowledge discovery and change.	18
Total		**75**

access and usage (30%), data publishing (26%), and data sharing (23%). Of relevance to efforts to design a more integrated data and computing infrastructure was the finding that most scientists access data and compute resources via web interfaces or remote login rather than application program interfaces (APIs) and therefore are currently not set up to flexibly leverage more integrated capabilities across the DOE complex.

The survey questions were divided into different categories. Out of these, knowledge gathering, managing, and sharing (KD) were identified as the overall area of greatest need, followed by human-computer interaction (HCI). The topics covered in these categories can be found in Table 3, p. 11.

Survey questions focusing on the effective use of exascale systems received mixed results, pointing to a potential need for further education of the wider community. For example, new techniques for working with deep memory hierarchies on extreme-scale computing systems reached only an average of 3.24, and the direct data delivery into ASCR computing systems from BER data resources faired only marginally better with a score of 3.86. On the other hand, ingestion and access to large volumes of scientific data garnered a score of 4.49/39%, and the petascale-related topic of *in situ* analysis of observational, experimental, and computational results achieved 4.40.

A list of questions ranked by average rating can be found in Appendix 3, p. 45.

Table 2. Top 10 Needs Identified by the Survey

Survey Question	Average Rating or Percentage in Highest Need Category
An easy way to publish and archive data using one of the DOE data centers	4.79
A means for comparing diverse data types generated from observation and simulation	4.71
User support for data access and usage	4.64
Access to sufficient observational and experimental resources	4.58
Access to enough computational and storage resources	4.52 / 41%
Method of ingesting and accessing large volumes of scientific data (e.g., from a data archive to supercomputer)	4.49 / 39%
Quality control algorithms for data	4.46 / 31%
A unified and single user account to access all BER and ASCR resources	4.44 / 38%
Reliability and resiliency of resources	34%
In situ analysis of observational, experimental, and computational results: the ability to interpret results and verify new insights within the context of existing scientific knowledge	4.40

Table 3. Top Needs Identified by Survey Respondents	
Survey Questions	**Average Rating**
KD: Method of ingesting and accessing large volumes of scientific data (e.g., from a data archive to supercomputer)	4.49
KD: Quality control algorithms for data	4.46
KD: Interfaces that ensure a high degree of interoperability for different formats and semantic levels among repositories and applications	4.18
KD: Capture of provenance information for data	4.11
KD: Reproducibility	4.06
HCI: Collaborative environments	4.31
HCI: Improved user interface design	4.00

4. Data Services Needed to Support Science Requirements

Workshop organizers held a session to explore in greater detail the use cases developed to exemplify the current scientific goals and challenges faced by the climate research community. In particular, participants were asked to identify the key data and computing challenges that the community encounters and the types of services that would have the most impact on their scientific discovery process. Participants were divided into two teams to discuss these topics, and both groups had similar responses.

Scientists identified the following data-related challenges in their research processes:

- **Data and Linked Resource Discovery** — Time-consuming searches for older data, potentially from papers, were identified as obstacles. Moreover, participants noted that data discovery alone was not sufficient because researchers also need to be able to identify related metadata, provenance, and tools to use the data with confidence and ease. Similarly, they need discovery methods with suitable computational and storage resources to analyze any identified data.
- **Multisource Data Treatment** — More solutions are needed for integration, correlation, and comparative analysis of data with different dimensionalities, geophysical properties, levels of quality, and related uncertainties. A particular challenge is the comparative analysis of observational and modeling data. There is a perceived lack of dialogue about data harmonization between the observation and modeling communities. Several workshop participants noted that efforts to improve connections between these two communities have been made in recent years by CESD, including its Atmospheric Radiation Measurement (ARM) Climate Research Facility, Atmospheric System Research, Regional and Global Climate Modeling, and Earth System Modeling programs, along with others around the world. More recently, CESD's Environmental System Science activity conducted a workshop on model-observation integration, modeling frameworks, data management, and scientific workflows (U.S. DOE 2015).
- **Handling of Large Data Volumes** — Analytical tasks use increasingly large data volumes from multiple geographically distributed resources. The community wants new approaches that enable the efficient analysis of these data sets without the need for massive data transfers.
- **Tool Flexibility** — Data tool challenges include easing adoption, scalability, and adaptability; determining future needs; and addressing issues with cross-tool integration and associated training and education. Many useful community tools were developed in a different era when data volumes were smaller and analysis processes were carried out on local systems with single processors. The community would like to continue leveraging the knowledge and capabilities encapsulated in these tools (e.g., trusted, community-wide, and standardized mathematical approaches) but in a more scalable environment with more modern user interfaces. Participants also sought advice in terms of good data models and approaches that could ease integration of tools into complex data analysis workflows.
- **Reproducibility** — Scientists need practical solutions to enable reproducibility of their work, be it modeling or data analysis tasks. Their focus is primarily on *scientific reproducibility* (replication of conclusions with different methods) and *computational reproducibility* (the same results with the same modeling setup).
- **Data Publication and Attribution** — Researchers seek guidance and support on standardized ways to publish data that integrate well with the community's journals and expectations. Furthermore, provisions are needed to ensure that all researchers have access to the required long-term storage and curation capabilities that would accompany these formal data publication efforts. A central discussion point was attribution, which must go hand in hand with the data publication effort. Data products often are

based on the work of many others. Moreover, data sets are integrated and refined at different phases of the research process, from raw data collected from heterogeneous data sources to the final publication of a data set used to validate a climate modeling campaign. Community-determined standards are needed regarding who should be cited at which step. The concern that data could be used inappropriately also was discussed, underscoring the need for methods that allow researchers to engage with others on subsequent use of their data.

Based on these identified challenges, the two teams discussed needed general data services. Solutions include:

- Publication of a notional data service architecture (e.g., a taxonomy of what data services are provided and where).
- Data services partitioned by size (downloadable versus very large).
- Discovery based on metadata that describe the conditions and context under which data were collected.
- Standardized data and metadata formats across observational and modeling data to enable easier integration and comparison.
- Server-side computations that push algorithms to the data rather than downloading the data.
- Intelligent data services that inform users of other related data products that may interest them (e.g., data recommendation engines).
- Persistent links to a specific data set that can be published or accessed in the future without repeating a complex search.
- Means of avoiding duplication of data downloads to community computing resources.
- Programmatic access to data services allowing their easy use in scientific workflows.
- Collaborative workspaces.

The four highest-priority requirements identified are server-side data subsetting and analysis; better data documentation; sufficient data and computing capacity, including dedicated resources for data science; and standardized interfaces between tools and infrastructure services.

Scientists also would like to find in such a data environment synthesized observational data products that support model development and evaluation including, for example, ARM Best Estimate products and Observations for Model Intercomparison products. These data sets should be accompanied by robust quality control algorithms and uncertainty quantification assessments and be linked to tools that support data merging, processing, and further analysis. Furthermore, they should be easily accessible and usable in model development test beds.

Research communities and data service providers see two key impediments to creating these types of services: lack of dialogue and coordination across disciplinary boundaries and insufficient funding for such efforts (i.e., a stable funding stream for long-term operations is needed). Should these key impediments be addressed, software developers highlighted a number of additional challenges such an effort would face. These included overcoming current requirements for multiple authentication and authorization layers, making sufficient computational resources available, and developing the necessary data and scientific expertise to enable all to participate in this new environment.

Scientists agreed that a successfully implemented infrastructure not only would accelerate scientific discovery processes through higher-performance tools and removal of redundancies, but also, more importantly, would enable new science and discoveries through easy experimentation with novel data analysis approaches.

5. Advanced Computational Environments and Data Analytics

Advanced computational environments supported by key climate modeling, observations, and data centers—such as the national scientific user facilities funded by the DOE Office of Advanced Scientific Computing Research (ASCR)—provide the DOE research community with a number of tools and services. Such capabilities include high-performance computing (HPC), clusters, robust short- and long-term storage, networking, and coordinated software resources and tools. These major ASCR computing facilities include the Argonne Leadership Computing Facility, National Energy Research Scientific Computing Center, and Oak Ridge Leadership Computing Facility. They currently have different architectures (e.g., graphics processing units versus accelerators), programming models, and operating environments (e.g., hardware, software, policies, security layers, and queue management) running on multiple systems. In addition to running and processing state-of-the-science climate information at these facilities, the CESD research community must rely on multiple levels of services to effectively manage, analyze, and visualize distributed data from many sources.

Moving from the present computational environment to a federated system of tools and services will require, among other tasks, ensuring that the following levels of services be robust, resilient, and consistent throughout (see Fig. 2, p. 17):

- **Common Data Services** — Includes data movement, curation, long-term preservation, discovery, exploration, and more. These services will be shared across all CESD projects and, hopefully, with other research communities as well.
- **Domain-Specific Distributed Data and Analytical Services** — Captures the set of unique requirements and services needed for each CESD climate project. These include, for example, software performance [e.g., parallel input/output (I/O), analysis, and data set transformation] and data analysis services with better I/O bandwidth and more memory for analyzing and computing ever-expanding data sets.

> Achieving community scientific goals requires additional storage and computing resources, along with a common virtual computational environment that conforms to established standards across DOE Office of Advanced Scientific Computing Research computing facilities.

- **Data Systems Software Layers** — Includes standardized lower layers of software services such as metadata, directory structures, provenance, extension of bit-level verification, and workflows that allow reliable and unlimited access to computational and analytical resources with well-defined, scriptable community APIs. Another avenue of services provides the ability to dependably archive and serve data where the user can adjust the cost, speed, and reliability of the underlying storage service.
- **Data System Computational and Storage Hardware** — Includes HPCs, clusters, clouds, and dedicated large-scale archives for modeling; *in situ* data analysis; and *post-hoc*, large-scale computational data analysis. This service also includes in-transit data processing to enable extreme-scale climate analysis and an emerging ability to provide highly reliable, geographically distributed storage (which should be further explored).
- **Networks** — Binds the collection of disparate hardware, other networks, and software resources for CESD community use. Networks are also necessary to replicate and move large data holdings at storage facilities and to federate connectivity. The 100 gigabit DOE Energy Sciences Network is of particular interest, along with facility implementation of data transfer nodes. Connections between the facilities and community imply improvements to Globus/GridFTP and data endpoints (e.g., disk-to-disk and disk-to-tape).
- **Portability** — Requires that flagship computing facilities' operating environments and methods

not be unique so that scientific workflows can be interchangeable among the centers. The Accelerated Climate Modeling for Energy project is one example where workflows must reliably operate the same across ASCR computing facilities.

- **Support** — Encompasses user support for reliable access to computational resources, data transfers, login access, persistent data preservation, stakeholder training and outreach, and general system use and documentation.

If CESD is to optimize its data investments—and thus the scientific impact of its observational and modeling programs—it must ensure that a common virtual computational environment is in place. Moreover, a significant fraction of that environment should be shared among the different activities of CESD and international communities, rather than having specific domain environments for each project. Therefore, an integral part of CESD's overall science strategy should include a comprehensive, long-term, and sustainable solution for empowering domain-specific distributed data services, data system software layers, next-generation HPC and storage, and next-generation networks that access large-scale national and international data sets. Community-established standards and protocols are needed for distributed data and service interoperability of independently developed data systems and services. A reference model and supporting API standards are essential for enabling collaborations and facilitating extensibility whereby similar, customized services can be developed across CESD science projects, as shown in Fig. 2, p. 17. The environment must support the ability of resources at every level of the figure to transfer information within and across the multiple layers of services.

To address usability issues, more comprehensive and constantly up-to-date documentation would exist to aid scientists in hardware, software, and infrastructure discoverability, availability, and access. Key hardware issues include storage, cores, memory, and compute interactions. Today, the use of hardware has a steep learning curve, with multiple levels of integral security details (e.g., credentials, authentication and authorization, tokens, and virtual private networks) and different resource and service restrictions for each compute facility. Managing and analyzing distributed data for petabyte archives consisting of 100-terabyte data sets necessitate both long-term storage for observations and short-term scratch space for large-scale computational experiments. Diversity of compute resources must be standardized across the facilities so that similar programming models (such as FLOPS-intensive versus data-reducing) are reliable, resilient, and, above all, consistent among the virtual facilities. Containerized performance-portable methodologies could be addressed by multilevel computing approaches with shared storage and an archival high-end, compute-intensive, mid-range, and data-intensive architecture and typical cluster resources. This approach also will include compatible I/O and memory performance for large-scale data sets. System usability should enable nonexpert users to accomplish large-scale data analysis and allow all users to simply navigate the batch queuing system.

If data are housed at a major facility or data center or distributed across many facilities, then moving large amounts of that data in a reasonable amount of time to compute facilities (for remote processing) or to data storage (for replication and backup) is feasible. This will allow data federation to be managed differently than the way researchers interface with data today (i.e., most users download data to their home organization for analysis and visualization). Once data have been created, produced, or reduced, the data need to be published or republished as a service so that it can be used by other members of the community without large-scale data movement. This approach makes remote or local data manipulation and publication available to all, including cloud services that will complete the full spectrum of data availability and accessibility.

From the resource providers' and software developers' perspectives, the primary impediment to computational environment and data analytics development is continuity of funding. Keeping up with heavy user demands and disruptive technologies for this type of environment will require sustained monetary resources. Therefore, a sustainable business model for CESD-wide data infrastructure and environments is warranted; cost justifications and metrics of success will be evaluated and determined in terms of scientific productivity enhancements. Additional key impediments include remote compute services and more short- and long-term storage (i.e., rotating and tape archives).

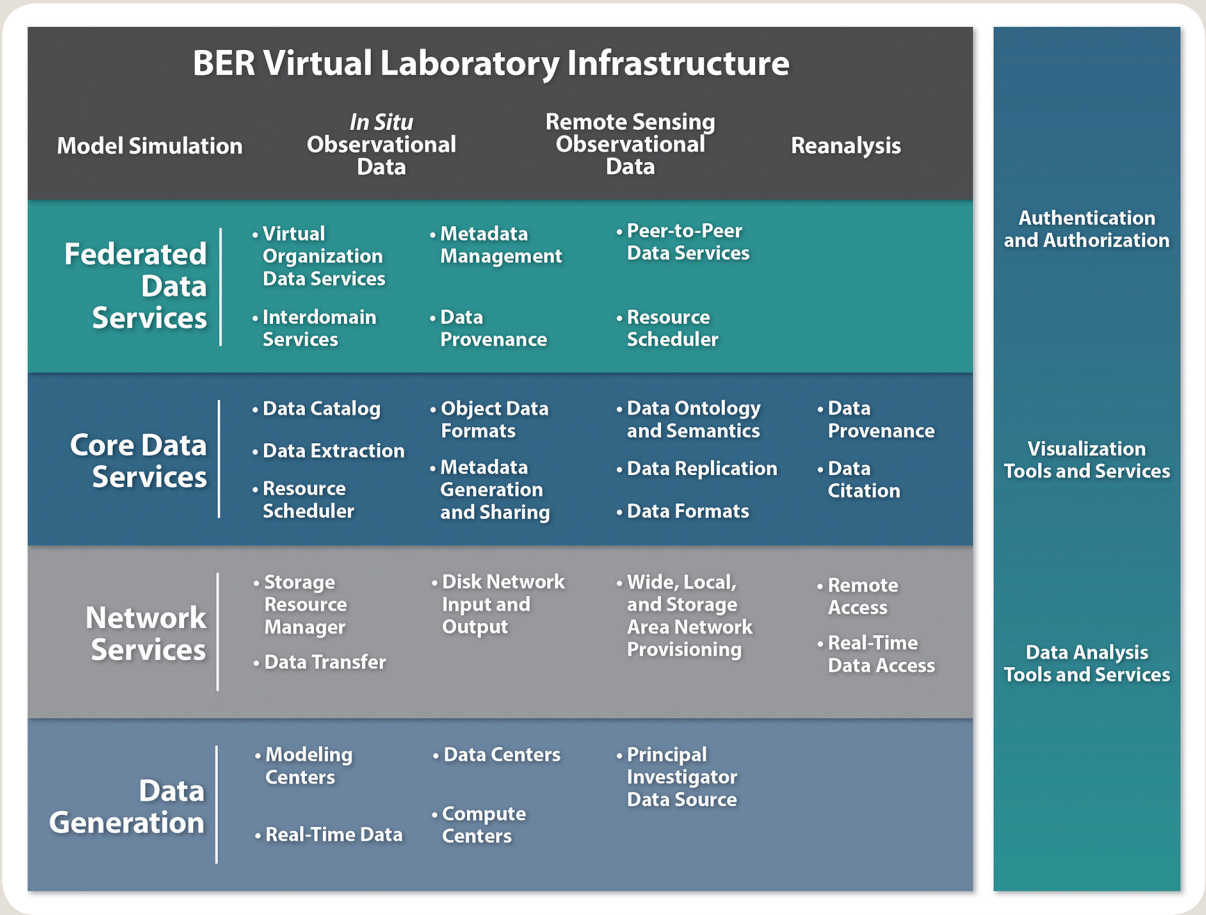

Fig. 2. Framework and relationships for distributed, federated climate data products and services that can support powerful, flexible, and advanced computational virtual environments and data analytics. Each hosted service will be exposed through a set of simple and well-documented web-service APIs (layered with security when appropriate) so that different kinds of clients can easily execute invocations and perhaps chain requests in complex scientific workflows.

The prioritized needs for the virtual federated computational environment include:

1. **Hardware** — More petabyte-scale storage is required, along with compute cores and memory for co-located data computing. Also necessary is coordination of hardware efforts with ASCR peta-scale and exascale HPCs. Workshop participants especially noted compatibility difficulties between compute core technologies and software, indicating that ever-changing code revisions are needed as technology shifts back and forth.

2. **Simulation and Observational Storage and Preservation Strategy** — This need centers on publishing data so that it is usable by other members of the community. Preliminary inference from the use cases indicates that the average CESD project publishes 500 terabytes of data a year. One or two CESD projects, such as the Coupled Model Intercomparison Project (CMIP), expect to publish tens of petabytes of data over the project's lifespan.

3. **Data Analysis, Retrieval, and Reduction** — Standardization of the analysis framework is needed, as well as fat nodes with high-throughput I/O and memory.

4. **Support** — Multiple classes of computational analyses must be supported with the federated environment.

5. **Documentation** — Requirements include up-to-date documentation detailing resource availability

and specific user guides for analysis packages. Providing users with training and access to white papers that outline next-generation computational environments also would be useful so that DOE science and CESD infrastructure can evolve in lock step and upcoming projects can fully leverage newly available resources.

6. **Operational Support** — Facility support for operational services and data archives (e.g., CMIP) is needed.

The virtual federated environment also must allow scientists to access and compare observational data sets from numerous sources including, for example, Earth Observing System satellites and the ARM sites. These observations, often collected and made available in real or near-real time, are typically stored in different formats and post-processed for conversion to a format allowing easy comparison with model output (i.e., CMIP). The need for providing both on-demand and value-added data products adds another dimension to the required capabilities. Finally, science results must be applied at multiple scales (e.g., global, regional, and local) and made available to different communities (e.g., scientists, policy-makers, instructors, farmers, and industry). However, providing results to the science community will take precedence over all other user communities.

6. Data Centers and Interoperable Services

Data centers supported by CESD handle diverse scientific data products, from multiple petabytes of climate model data to field and experimental data. These centers use a variety of tools and technologies to manage and share their data. Some data centers also provide interoperable data and services to broader scientific communities including, for example, Observations for Model Intercomparisons, the Thematic Real-time Environmental Distributed Data Services (THREDDS) data catalog, and International Organization for Standardization (ISO)-19915 metadata standards. Fig. 3, this page, illustrates an integrated cyberinfrastructure using various interoperable services.

Workshop participants discussed the data centers supported by CESD and other agencies and examined their current interoperable services. Key points from the discussion follow.

> Identifying, applying, and following key interoperability enablers are all critically important when developing tools for CESD programs and projects. Such enablers include metadata conventions and standards, workflow and provenance capture, and data and visualization protocols.

6.1 Earth System Grid Federation

ESGF, one of the largest-ever collaborative data efforts in climate science, is now used to disseminate model, observational, and reanalysis data for research assessments and model validation (see Fig. 4, p. 20). ESGF is an international multiagency-driven activity led by

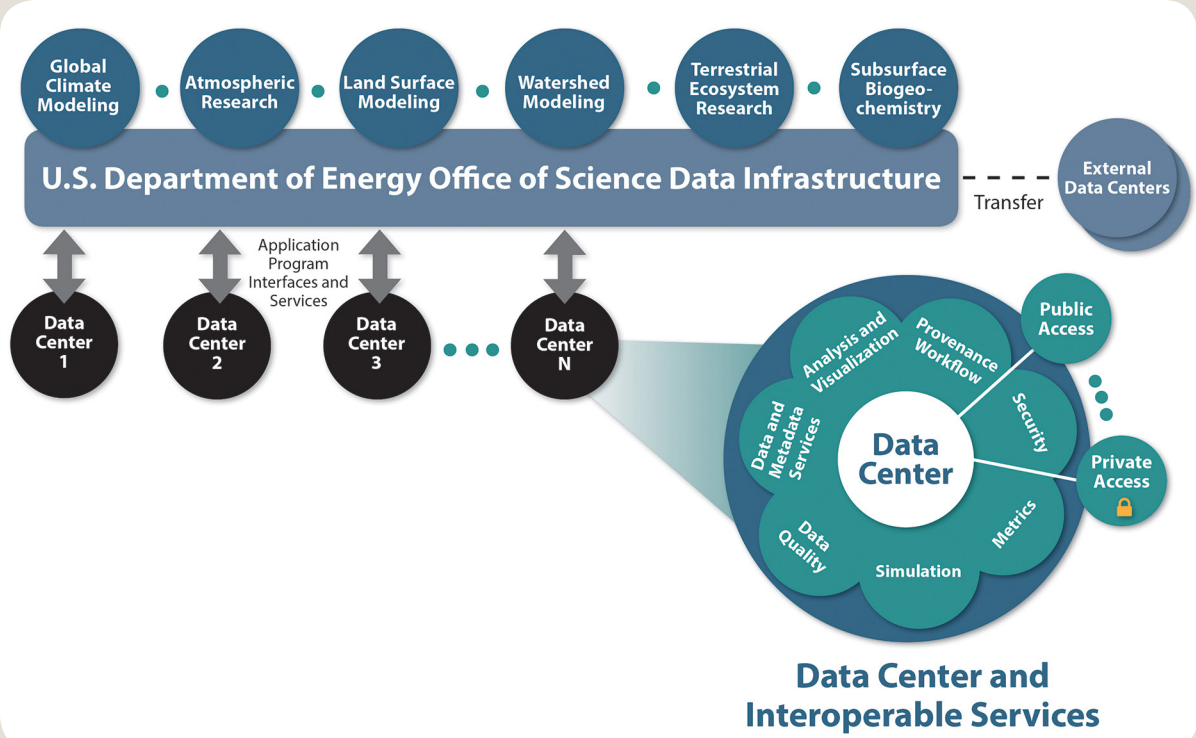

Fig. 3. A concept diagram of an integrated cyberinfrastructure leveraging core DOE Office of Science resources to enable discovery, analytics, simulation, and knowledge innovation. [Modified from Williams, D. N., et al. 2014. "Department of Energy Strategic Roadmap for Earth System Science Data Integration." In *Proceedings of the 2014 IEEE International Conference on Big Data*. Washington, D.C., pp. 772–77. ©2014 IEEE]

DOE as an open-source, operational code base with secure, petabyte-level data storage and dissemination of the resources essential for studying climate change on a global scale. ESGF is designed to remain robust even as data volumes grow exponentially. Virtually all climate science researchers in the world use it to discover, access, and compute data. ESGF's decentralized approach has changed relatively recently from a client-server model to a more robust peer-to-peer approach already proven for distributing large amounts of data and information. A system of geographically distributed peer nodes comprises ESGF. These nodes are independently administered yet united by common protocols and interfaces, allowing access to global atmospheric, land, ocean, and sea-ice data generated by satellite and *in situ* observations and complex computer simulations for use in national and international assessment reports. Scientists are accessing climate data more efficiently and robustly through newly developed user interfaces, distributed or local search protocols, federated security, server-side analysis tools, and other community standards—all for improving the understanding of climate change.

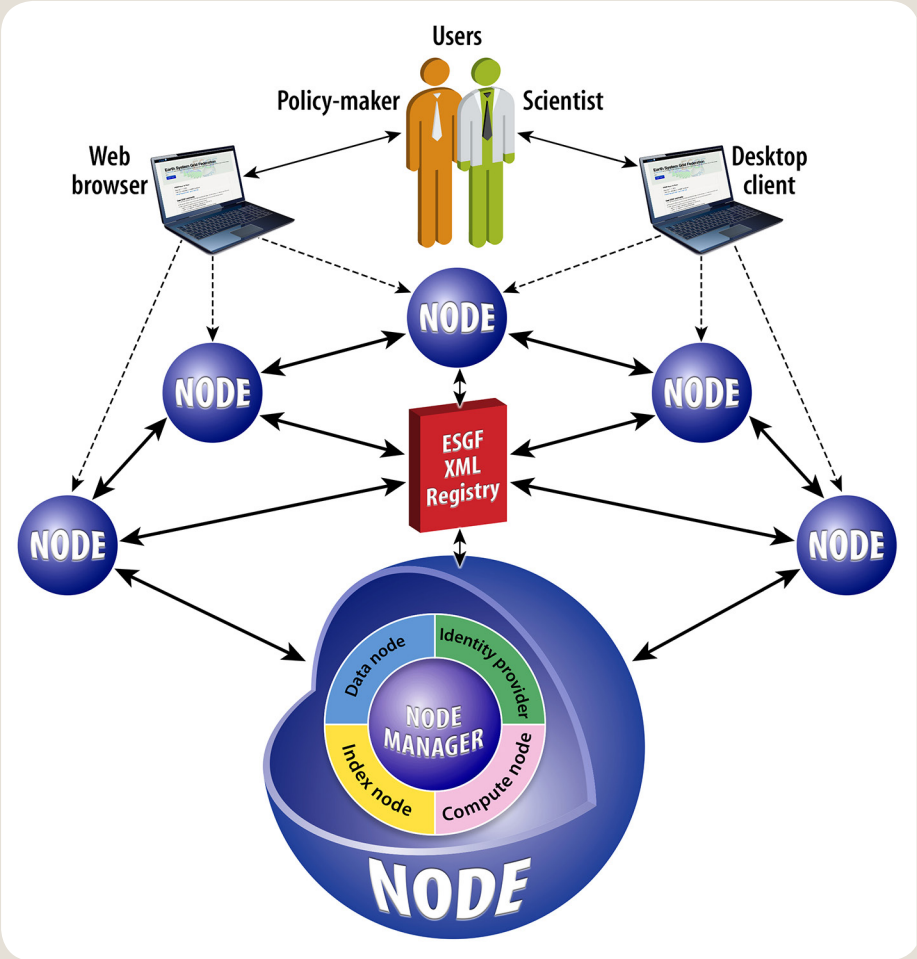

Fig. 4. The Earth System Grid Federation (ESGF) ensures equal access to large disparate data sets (e.g., simulation, observation, and reanalysis) that in the past would have been accessible across the climate science community only with great difficulty. The ESGF infrastructure enables scientists to evaluate models, understand their differences, and explore the impacts of climate change through a common interface, regardless of data location.

ESGF's architecture can easily be leveraged for accessing data from other scientific domains, such as satellite, instrument, and other forms of observational data. ESGF is now in the early stages of being adapted for use with the National Aeronautics and Space Administration's (NASA) Distributed Active Archive Centers (DAACs), published data archives of the National Oceanic and Atmospheric Administration's (NOAA) National Centers for Environmental Information, and international research communities' data exchanges. The importance of ESGF continues to grow as computing platforms and archives expand and reach extraordinary speeds and capacity.

6.2 ARM Climate Research Facility Data Center

The Atmospheric Radiation Measurement (ARM) Climate Research Facility operates field research sites around the world for global change research. Three primary locations—the Southern Great Plains megasite, North Slope of Alaska megasite, and Eastern North Atlantic site in the Azores—are heavily instrumented to collect massive amounts of atmospheric data, as are ARM aircraft and the portable ARM Mobile Facilities. ARM data are freely available to the scientific community in near-real time. As part of this effort, ARM scientists and infrastructure staff provide value-added processing to data files to create new data streams called value-added products that apply scientific algorithms to convert instrument-measured variables to geophysical variables or that combine observations from multiple instruments into a single data stream. In addition, the ARM Data Center archives and distributes data products contributed by principal investigators (PI) and from field campaigns.

The ARM Adaptive Architecture (see Fig. 5, this page) is being developed to provide data tools, connections, and software for scalable microservices to support diverse observational data sets. Many interoperable services such as machine-readable data quality; data-flow monitoring; next-generation data discovery; and data visualization, extraction, and analysis capabilities will be delivered through tools such as:

- ARM Data Integrator,
- Python ARM Radar Toolkit (Py-ART)
- Data System Status Viewer
- Data Delivery Tracking
- PI data product registration (Online Metadata Editor)
- Data Discovery portal

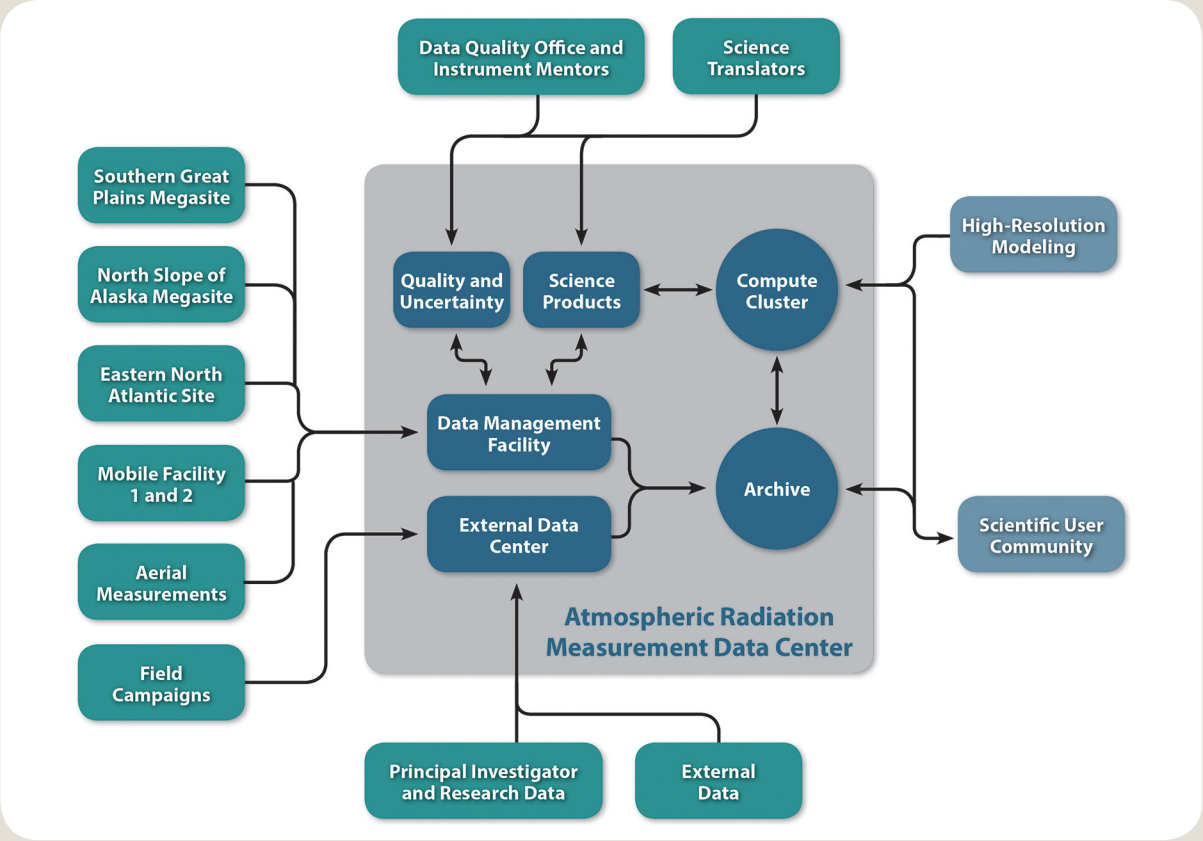

Fig. 5. The ARM Adaptive Architecture is being specified and evolved to provide data tools, connections, and software for scalable microservices.

- Data citation tool using automated digital object identifier (DOI) generations
- THREDDS
- Big Data analytics using No-Structured Query Language (e.g., Cassendra and Hadoop)

6.3 Carbon Dioxide Information Analysis Center

CDIAC offers publicly available data and value-added products for climate change research. CDIAC's data collection is diverse, reflecting the breadth of climate change research, and includes atmospheric, oceanic, terrestrial, climatic, and anthropogenic emissions holdings. Considerable effort is devoted to the development and production of science-driven global- and regional-scale synthesis products, such as AmeriFlux, the Global Ocean Data Analysis Project, and estimates of global and national fossil-fuel carbon dioxide emissions. CDIAC hosts and serves processed data from measurement networks (e.g., the Advanced Global Atmospheric Gases Experiment and the Total Carbon Column Observing Network), intensive field campaigns (e.g., Next-Generation Ecosystem Experiments–Arctic, Spruce and Peatland Responses Under Climatic and Environmental Change, HIAPER Pole-to-Pole Observations), and other projects (e.g., Global Carbon Project). A searchable catalog based on standards-compliant metadata enables easy data discovery, and customized interfaces allow users to query, visualize, subset, and download many CDIAC collections. Multiple data formats are offered for most data holdings to facilitate broad use.

CDIAC is evolving from an independent data center to an integral part of a federated data system that includes the ESGF and NASA DAACs (see Fig. 6, this page). As part of this federated system, CDIAC will develop data tools and services to facilitate interdisciplinary research across multiple data holdings and scales and will benefit from existing tools and future developments

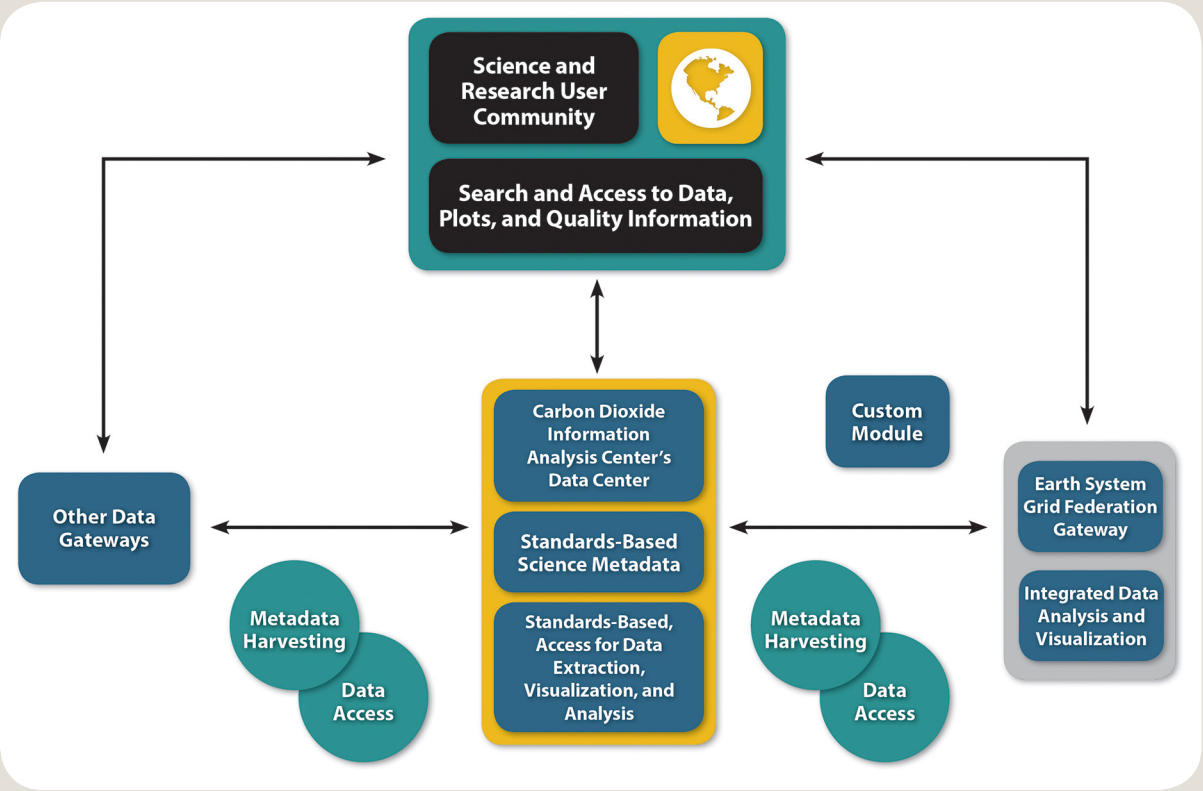

Fig. 6. Diagram illustrating publication of CDIAC data and metadata to the Earth System Grid Federation.

elsewhere. Existing workflows, processing capabilities, and automation will be leveraged and expanded to support current and future research within CESD's Environmental System Science activity.

6.4 Other Interoperable Services

Globus (www.globus.org) provides high-performance, secure, and reliable data transfer, sharing, synchronization, and publication services for the science community. With its user-friendly web interfaces and simple application program interfaces (APIs), Globus is easily integrated into services such as ESGF, the National Center for Atmospheric Research's (NCAR) Research Data Archive, and the DOE Systems Biology Knowledgebase.

Other BER data centers and user facilities, such as the Environmental Molecular Sciences Laboratory, and their interoperable services will be included in future discussions.

The Compute and Data Environment for Science (CADES) infrastructure at Oak Ridge National Laboratory offers hardware hosting with the ability to deploy custom software stacks to meet diverse user needs and also should be considered in future services discussions.

In addition, workshop participants discussed various data services from external data centers such as NASA's Giovanni reanalysis and regridding service, its Open Geospatial Consortium and metadata services offered by DAACs (10 data centers), and satellite data services provided by NOAA's National Centers for Environmental Information (formerly the National Climatic Data Center) snow and ice center. In addition, the group discussed the megaportal services offered by the National Science Foundation's data-stewardship engineering team whose members include staff from Unidata, the University Corporation for Atmospheric Research, and NCAR.

6.5 Recommended Interoperable Services

Workshop participants prepared the following list of required and recommended interoperable services to communicate between centers.

- Server-side analysis and visualization (analysis as a service versus downloads) should be scalable, robust, resilient, easy, and tested. These services also should allow users to cache recent analyses in a sharable way (re-use an analysis where possible) and to isolate function from implementation [an analysis should specify what is done, not on which machines (i.e., say "no" to shell scripts)].
- Common metadata across data sets should be based on properties, features, and temporal-geospatial variables and should include provenance and versioning for reproducibility (see Fig. 2, p. 17).
- Seamless unified search and access across data sets are critical components for enabling interoperability. These search capabilities should provide common indices, hashes, duplicate detection, and quality control information, methods, and data where possible. They should allow API-based access to data sets, data services and catalogs, measured reuse of data, citation, and acknowledgments.
- Data services preferably should be built based on open-source software licensing. Curation and comparison of stewardship policies across centers should be considered, involving an index of policies informing data management plans, data persistence, number of copies, and policy best practices. Other considerations for enabling interoperable services are high-speed data transfer services to support large-scale data analyses built using current best practices, such as a large-scale data movement infrastructure using the Science DMZ model (Dart et al. 2013) and Globus services (Chard, Tuecke, and Foster 2014).

7. Inventory of Existing CESD Computer Resources, Data Tools, and Services

CESD data projects use a variety of data management tools and technologies, many of which are open source, community developed, and used by multiple projects. The data tools required by these projects are diverse and span a wide array of needed capabilities. Currently, no one tool can handle all of CESD's diverse data needs, which presently are not being completely met, despite the wide array of available tools. Gaps still exist, for example, in areas such as quality assurance and control, interaction with gridded data, and metrics.

The desired goal is a healthy, sustainable ecosystem of tools that together serve the diverse data needs across projects. The first step toward meeting this goal is to develop an inventory of existing data management tools used within CESD projects (see Table 4, p. 26). Next, benchmark testing of existing tools should be conducted to evaluate their potential for broad adoption within the Virtual Laboratory infrastructure. In addition, standardization of storage formats, APIs, authentication and authorization, and identifiers can significantly improve tool interoperability while enabling a healthy competition among available tools.

Workshop participants highlighted as key data capabilities the need for seamless, unified search and access across data sets; uncertainty quantification tools; and connection to a specific workflow. Server-side analysis and visualization, single sign-on and federated authentication, and tools to combine disparate data sets at different resolutions also were identified as important. These server-side analytics should consider the total costs of data movement and analytics ease for users. Related needs include flexible and scalable virtualized approaches that allow growth of the analytics over time. Virtualized and container-based approaches can enable new analytics functionalities to be systematically added over time. Furthermore, provenance tools such as VisTrails need to be integrated with project-specific workflows.

Suggested action items include building an inventory of tools used by major projects, developing a strategy to integrate tools and services across facilities and infrastructures, providing tools as a service in the computing architecture, enabling a source code repository that is "common" with front-end-release via web browsers, and providing precreated virtual machines/Red Hat Package Managers with a representative set of tools. Participants noted that structuring these needs and requirements in an actionable manner for computing and observational facilities is essential to success.

> An inventory is needed outlining the available data, compute tools, and resources currently used by CESD and its associated research communities. Evaluation and assessment of these shared data, tools, and resources would ease their route to adoption into the integrated data ecosystem.

Also detailed were some potential methods for assessing tool maturity and capabilities. Suggestions include:

- An app store–style star rating or clearinghouse.
- Publication references (digital object identifiers for tools like Zenodo).
- Metrics tracking (e.g., most recent activity and the number of contributors, diverse scientific projects that the tool supports, downloads, users, and usage).
- Assessment of the commitment level of developers to sustainability and software engineers to support.

Participants also discussed other action items related to existing tools and services benchmarking, such as software maintenance, security patches, connectivity to high-performance computing resources, maintainability, installation, and documentation. A related topic was the types of support that the science community expects. These expectations were diverse and included software documentation and maintenance; user

Table 4. Open-Source Tools that Should Gain Wider Accessibility Within the CESD Community

Tool	Need
Infrastructure	Use flexible, extensible infrastructure tools for future CESD efforts and partnering DOE projects to automate laborious, repetitive simulation data tasks and to heighten productivity and user experience. The same infrastructure must allow CESD scientists to access and compare data sets from multiple sources (e.g., simulations, reanalysis, and observational satellites and instruments).
Globus transfer, sharing, publication; GridFTP	
PERFormance focused Service Oriented Network monitoring ARchitecture (perfSONAR)	
Panda Global (data and job placement across facilities)	
Earth System Grid Federation	
Velo	
Docker	
perfSONAR for network data transformation	
Atmospheric Radiation Measurement (ARM) Data Integrator	
Python ARM Radar Toolkit (Py-ART)	
Multipurpose serial/parallel tools: NCAR Command Language (NCL), NetCDF Operators (NCO), climate data operators (CDO), Ultrascale Visualization–Climate Data Analysis Tools (UV-CDAT) (need scalable versions)	
Climate Model Output Rewriter [generates and checks for climate forecast (CF) metadata standards]	
International Land Model Benchmarking project (ILAMB), Earth System Modeling Framework regridder	
Toolkit for Extreme Climate Analysis (TECA) and Illiad for analyzing Hierarchical Data Format version 5 (HDF5) atmospheric data (high performance)	
Metadata	Metadata tools to discover, facilitate, and navigate the CESD data infrastructure.
Online Metadata Editor	
Mercury, Earth System Documentation	
Open-source Project for a Network Data Access Protocol, Thematic Real-time Environmental Distributed Data Services (THREDDS)	
User Metrics and Usage Analysis	Service-specific metrics to measure the usage and adoption of specific capabilities.
Data Quality and Instrument Monitoring	Tools to ensure data completeness and integrity for trusted use consumption.
Machine-readable data quality reports	
Instrument monitoring tools	

support; and support for data quality issues, provenance communications, community-used tools, and deployment.

Finally, the group tackled the topic of data and metadata conventions and whether they should be adopted across many or all data centers. Specifically discussed were climate forecast (CF) metadata conventions for model data and CF-type conventions for observations and experiments. Participants agreed that common metadata, provenance, and DOI standards (including common assignment and collection approaches) should be developed and that other

Table 4. Open-Source Tools that Should Gain Wider Accessibility Within the CESD Community

Tool	Need
Data Analysis and Visualization	Analysis framework that includes visualization information techniques and automated data manipulations, such as data mining, feature tracking, and reduction. Server-side and *in situ* computation is necessary as increases in data size and algorithm complexity lead to data- and compute-intensive challenges for CESD diagnostics, uncertainty quantification, analysis, model metrics, and visualization.
UV-CDAT	
NCL; NCO; CDO; Matlab; Interactive Data Language (IDL); Visualization and Analysis Platform for Ocean, Atmosphere, and Solar Researchers; R	
Open Source: R, Exploratory Data Analysis Environment (some versions), Py-ART	
Commercial: Matlab, IDL	
TECA (feature tracking)	
Accelerated Climate Modeling for Energy (ACME) diagnostics, Program for Climate Model Diagnosis and Intercomparison (PCMDI) metrics, ILAMB	
Uncertainty quantification: Dakota, Problem Solving environment for Uncertainty Analysis and Design Exploration (PSUADE)	
Collaboration and Work Management	Tools to speed up, track, manage, and monitor key tasks, software, and infrastructure resources.
Confluence, JIRA, ServiceNow, Git, Pegasus	
Wiki	
NX technology to work on remote machines	
Citations and Publications	Unique data and user identifiers that link to data and metadata.
DOI tools (DOE Office of Scientific and Technical Information), Open Researcher and Contributor Identifiers (ORCID), Globus publication service	
Compute and Storage Facilities	High-performance computing (HPC) facilities deploy HPC systems, high-end storage, and data transfer nodes designed for accelerated scientific discovery.
Argonne Leadership Computing Facility, National Energy Research Scientific Computing Center, Oak Ridge Leadership Computing Facility	
Portal and Search Systems	Web-based user interfaces and content management systems for interactive tools and infrastructure use.
CoG, Drupal, WordPress D3, Solr, Elastic	
Workflow and Provenance	Implemented APIs to capture workflow progress and provenance in infrastructure.
Swift, Tigres, Akuna, VisTrails, ProvEn, Jupyter	

community-followed standards such as Hierarchical Data Format version 5 (HDF5), comma-separated values, and International Organization for Standardization (ISO) also should be supported for broader data integration.

8. Data Services and Monitoring

Given that participants identified increased reliability and resiliency of resources as a top-level requirement in the online survey (see Section 3, Survey Results, p. 9), data monitoring and computing and networking service needs are particularly significant within CESD's proposed integrated infrastructure. Subsequent workshop discussions of the survey results supported the idea that scientists perceive the performance of existing resources as unreliable, especially when used as part of more complex work processes across several resource types and institutions. However, exchanges in this broad group of workshop participants demonstrated that users do not want to get involved in the operational aspects of the resources. Rather, they expect the facilities to provide easy-to-use, reliable services and identify and resolve issues proactively.

For their part, service providers identified a range of challenges that occur when supporting users in a distributed environment. Foremost is the challenge of exchanging comparable monitoring information across facilities. The use of software-as-a-service (SaaS) services such as Globus has been shown to improve overall system reliability by providing a robust, centralized location for problem detection, determination, and correction. PerfSONAR provides a network layer example of how such information sharing can help with early identification of potential problems and their solution or mitigation (e.g., transmit via a different route or temporarily store data in a different place). However, this approach requires that service providers operate compatible monitoring services that capture similar information, as well as the ability to connect and evaluate overall infrastructure health. Users and service providers also identified the need for an event alert system that informs them of infrastructure issues at different levels of detail. Participants suggested that a designated CESD working group investigate solutions developed by the Large Hadron Collider collaboration to manage its worldwide network of resources.

Infrastructure users, in turn, suggested completely new types of monitoring services to include in an integrated CESD infrastructure. These services would focus on capturing metrics on users, data downloads, feedback on downloaded data, and publications resulting from the reuse of a user's data by others. In addition to data, users would like similar services for the software tools shared throughout the infrastructure. Results of such metrics-capturing services should be available to both the data owners and software developers and the data and software users. Discussions centered on technologies and approaches that would support the tracking of data as it is analyzed and combined with other data products, capturing not just bytes but also data reuse, impact, and attribution. Once again, SaaS approaches have much to offer in this regard. Of particular interest are inclusion of DOIs as part of downloaded data products and automated insertion of acknowledgement sections.

> A new class of monitoring services for the next generation of complex workflows would be valuable, particularly services that capture metrics on data and software downloads, users, and publications resulting from the reuse of a researcher's data and software by others.

9. Synergies with Peta- and Exascale Computing Hardware

In addition to local computing resources, climate and computational scientists are supported by DOE Office of Advanced Scientific Computing Research (ASCR) national user facilities, including the Argonne Leadership Computing Facility, National Energy Research Scientific Computing Center, and Oak Ridge Leadership Computing Facility. These facilities deliver a balanced high-performance computing (HPC) environment with constantly evolving hardware resources and a wealth of HPC expertise in porting, running, and tuning real-world, large-scale applications. HPC facilities currently deliver multiple petaflops of compute power, massive shared parallel file systems, powerful data analysis and visualization platforms, and archival storage capable of storing many petabytes of data. A transition to exascale computing will result in energy-efficient architectures with higher core counts and advanced data fabrics based on hierarchical memory technologies such as non-volatile random access memory. Data and flexibility-focused infrastructures—such as Oak Ridge National Laboratory's Compute and Data Environment for Science (CADES) and Argonne National Laboratory's Petrel and Magellan—when combined with ASCR HPC resources, offer opportunities for leading-edge techniques in data manipulation, storage, and end-user usability.

Synergy between CESD and peta- and exascale trends will hinge on leveraging technological advancement while maintaining a balanced computing environment that can support key collaborations among data infrastructure developers and HPC facility experts on the creation, debugging, production use, and performance monitoring of HPC parallel applications. The computing requirements of the CESD community already are tightly integrated into plans for future systems, and continued dialogue can maintain those synergies.

Current major HPC facilities include petaflop systems featuring varied and disruptive HPC technologies, along with Lustre and general parallel file systems capable of storing petabytes of data. The computing infrastructure includes heterogeneous underlying hardware and software and cloud platforms to meet user needs and employs large multicore, multisocket Linux clusters with a variety of processor types including graphics processing units. Partnering across the DOE Office of Science and National Nuclear Security Administration laboratories, the HPC facilities are preparing to launch several pre-exascale HPC systems set to bring hundreds of petaflops of computing power to the scientific community.

> DOE Office of Advanced Scientific Computing Research (ASCR) facilities need a policy for retaining data sets with a useful lifespan that extends beyond supported compute facility programs [e.g., the Innovative and Novel Computational Impact on Theory and Experiment (INCITE) program]. Establishing a single sign-on for authentication and federated access also would ease researchers' use of multiple ASCR computing hardware and resources.

In the past decade, HPC facilities have deployed many Linux clusters containing thousands of nodes. Most clusters have similar commodity node-based architectures and provide a common programming model for ease of use. That is, they are built and maintained using commodity off-the-shelf hardware and open-source software. Node components are selected for performance, usability, manageability, and reliability. Most Linux clusters at DOE facilities run a common software environment based on Red Hat Linux with added kernel modifications, cluster system management, monitoring and failure detection, resource management, authentication and access control, development environment, and parallel file systems. Many of these components are developed and maintained in house; others are developed and maintained in collaboration with a vendor partner. In the future, more of the existing cluster-scale technologies will migrate into the compute node itself, and renewed attention will be given to the interconnects and memory hierarchies that will comprise exascale systems.

HPC and other DOE facilities deploy dedicated data transfer systems, called data transfer nodes (DTNs), for moving data among facilities as required by science teams. In most cases, DTNs are deployed in Science DMZ environments, which enable high-speed connectivity among DTNs at different facilities and research institutions by means of the 100 gigabit per second Energy Sciences Network (ESnet). The DTNs run Globus software for convenient access to high-speed data movement capabilities. To develop large-scale and reliable disk-to-disk data transfers, collaborations between HPC facilities and ESnet (along with other network organizations) are working to instrument the hardware with application software, such as with the Earth System Grid Federation (ESGF) and Globus. This software will allow for isolated sandboxes and workflow substrates for experiments and different scientific workflows.

The hardware system effort also combines traditional HPC with emerging cloud technologies. More specifically, these platforms use (1) virtualized high-speed InfiniBand networks, (2) a combination of high-performance file systems and object storage, (3) diverse analytics infrastructures including graph engines and memory-intensive computing platforms, and (4) virtual system environments tailored for data-intensive science applications. More emphasis is being placed on configuring these analytic environments to be cognizant of data-analytics application needs. For example, systems are increasingly configured so that the memory and storage hierarchy can perform data-proximal processing. Surrounding the data storage is a cloud of HPC resources with many processing cores and large memory coupled to the storage through high-speed network backplanes. Virtual systems can be tailored to a specific scientist and provisioned on the compute resources with extremely high speed network connectivity to storage and other virtual systems.

Finally, in addition to large-scale data analysis, systems are being used to host large-scale data services, such as ESGF node services at locations around the globe. The data stored within the federated infrastructure include simulation, observational, and reanalysis data for multiple intercomparison projects. Table 5, this page, provides examples of these capabilities for a limited-scale deployment with a constrained scope. These resources would be considered compute and storage building blocks for larger analytics needs and can be scaled up according to program needs.

Table 5. Example Capabilities, Descriptions, and Configurations of the Hardware System Components of HPC Facilities

Capability and Description	Sample Analytics Configuration
Persistent Data Services. Virtual machines or containers deployed for web services. Examples include Earth System Grid Federation, Global Data Services, Thematic Real-time Environmental Distributed Data Services (THREDDS), and file transfer protocol.	8 nodes with 128 GB of RAM, 10 GbE, and Fourteen Data Rate InfiniBand (FDR IB)
Database. High available database nodes with solid-state disk (SSD).	2 nodes with 128 GB of RAM, 3.2 TB of SSD, 10 GbE, and FDR IB
Remote Visualization. Enables server-side graphical processing and rendering of data.	4 nodes with 128 GB of RAM, 10 GbE, FDR IB, and graphics processing units
High-Performance Compute. Several thousand cores coupled via high-speed InfiniBand networks for elastic or itinerant computing requirements.	~100 nodes with 32 to 64 GB of RAM and FDR IB
High-Speed and High-Capacity Storage. Petabytes of storage accessible to all the above capabilities over the high-speed InfiniBand network.	Several storage nodes configured to support petabytes of RAW spinning disk and object store capacity
Long-Term and Persistent Tape Storage. Tens of petabytes of long-term storage accessible upon request. Data are staged to disk cache, and user is notified when requested data are retrieved.	50 PB (or more) of high-performance storage system tape archive
Geographically Distributed High-Speed and High-Capacity Storage. Many petabytes of high-reliability storage distributed across physical locations allowing for irreplaceable and high-value data to be stored more cost-effectively.	10 PB (or more) of high-reliability storage per site across several sites

10. Network Services

High-speed network services will enable fast, robust connections among participating DOE national laboratories and computing facilities, NASA, NOAA, the National Science Foundation (NSF), and international federated data centers, effectively transporting hundreds of petabytes of large-scale simulation and observational data. As an example, collaborating centers use GridFTP for data replication and backup, driven by Globus. These network services also use the national and international 100 GBps Internet connections provided by Energy Sciences Network (ESnet), Internet2, and other domain-specific networks.

The International Climate Network Working Group (ICNWG), an Earth System Grid Federation working group, is engaged in efforts to improve and sustain data replication and data transfer performance among major climate data centers (see Fig. 7, this page). The ultimate goal of this effort is to achieve managed, sustained disk-to-disk throughput of multipetabyte data sets among the centers for replication. Achieving this capability also will allow the CESD virtual data infrastructure to meet the heavy demands of moving large-scale data to centers for critically important compute operations such as federated uncertainty quantification calculations and ensembles.

> Advances in current high-speed reliable data movement are necessary for sufficiently meeting CESD data resiliency and backup needs.

With the advent of software-defined networking, a rich set of application program interfaces (APIs) for interacting with the network (such as setup and route direction) is possible. The data grid can program the switches to use disjoint routes when doing multistream large-data transfers for replication and/or federated computing.

For network performance measuring, perfSONAR could be integrated into the infrastructure. PerfSONAR measures network performance capabilities at end sites by using tools including Bandwidth Test Controller (for throughput testing every few hours) and One-Way Active Management Protocol (for low-bandwidth, one-way delay measurement and packet loss testing), which

Fig. 7. International Climate Network Working Group (ICNWG, icnwg.llnl.gov).

runs continuously. Results could be stored on a server, which can be viewed using an API or web browser.

To monitor network performance and services, a perfSONAR node (a virtual machine) must be deployed alongside participating standard nodes as representatives of that host environment. To maximize network services, a number of perfSONAR boxes will be installed within the infrastructure spanning the federated data centers and network domains. This will immediately help address and troubleshoot local-area and wide-area network issues.

11. Participation with Broad Multiagency Data Initiatives

As DOE considers the design and implementation of a broad capability for data and informatics to support its climate and environmental science missions, it is imperative to catalog existing and emerging capabilities across multiple institutions and agencies, including international efforts, and determine how best to integrate new and existing capabilities. The development of a robust predictive understanding of Earth's climate and environmental systems is an inherently interdisciplinary problem. Integrating observational and experimental data, process knowledge, and predictive modeling across a wide range of traditional science domains, including physical, biological, and sociological, is necessary for development of sustainable solutions to pressing energy and environmental challenges. As DOE pushes forward to fully engage these challenges, a broad perspective on current and emerging data and informatics systems and their capabilities will provide the best opportunity for deep collaboration and rapid progress toward a system that serves agency needs while improving Earth science understanding for the whole community.

Major Earth science data and informatics systems and services are already operational in other U.S. agencies, including large-scale efforts at NASA, NOAA, and NSF. Those efforts are summarized in the sections that follow as an illustration of the depth and scope of current and emerging efforts in this area. A much more technical review of these and other existing programs is a critical first step in advancing DOE's capability in data and informatics systems for climate and environment. A wide variety of tools and technologies are being used, many of which are well evolved and could benefit a DOE system. Significant capabilities developed with DOE support also are available, as described in Sections 5–9, beginning on p. 15.

11.1 NASA

The Earth Observing System Data and Information System (EOSDIS) provides end-to-end capabilities for NASA Earth science data from multiple sources, including satellites, aircraft, and field measurements (earthdata.nasa.gov). The Earth Science Data and Information System (ESDIS) project manages the science systems of EOSDIS, providing science data to a wide community of users for NASA's Science Mission Directorate. Major ESDIS capabilities and objectives include: (1) processing, archiving, and distributing satellite data; (2) providing tools for archiving, processing, and distributing a variety of Earth science data; (3) ensuring ready access to data promoting research in the areas of climate and environmental change, guided in part by the gathering and analysis of data user metrics; and (4) promoting interdisciplinary data use.

> Strengthening partnerships with other national and international agencies is necessary for research community success.

ESDIS supports 12 Distributed Active Archive Centers (DAACs, earthdata.nasa.gov/about/daacs/), as well as many Science Investigator-led Processing Systems (SIPSs). A general view of ESDIS data flow is from primary instrumentation to a dedicated SIPS, where raw instrument data are processed to produce Earth Observing System (EOS) standard products. From SIPS, data move to the relevant DAAC for distributing, archiving, and use in a broad range of user services, including some value-added product generation and web-based access and analysis tools. Given the diversity of raw data sources (satellites, aircrafts, and field measurements) and science domains among the various SIPSs and DAACs, a coordinated strategy for documented interfaces has been an essential element in smooth operation of ESDIS. Design documents and interface control documents with standardized formats and information content are in place at each step in the data lifecycle—from instrument to DAAC and then to science users.

The EOSDIS data strategy includes a unified approach to gathering, indexing, and accessing metadata across

all its products, investigator-led teams, and data centers. The emerging metadata framework within EOSDIS is called the Common Metadata Repository (CMR), which brings together previously developed capabilities from the Global Change Master Directory (GCMD) and the EOS Clearing House. CMR will validate metadata adhering to various standards (e.g., ISO-19115 and GCMD's Directory Interchange Format) against a common standard (the Unified Metadata Model).

With numerous data centers (DAACs) and upstream service providers (including SIPSs), the integrated ability to search across the entire EOSDIS holdings is a crucial performance metric for ESDIS. The Earthdata Search application provides flexible keyword searching as well as a range of data discovery tools and services. Tools include web clients for browsing and ordering of mapped data sets, including time-varying data and open-source geospatial analysis tools (e.g., region-of-interest subsetting, reprojection, and geolocation). Another recently developed tool is the Global Imagery Browse Service, which helps solve the problem of many data sets being delivered in small "granules" that must be stitched together in space and time before arriving at first-look evaluations. ESDIS also provides a system for serving large and complex data sets to a broad range of users for near-real time applications [e.g., the Land, Atmosphere, Near real time Capability for EOS (LANCE)].

The collection of discipline-oriented DAACs is designed and operated as a distributed data and informatics system, with coordination managed through well-defined interfaces and standards. A special ESDIS Standards Office provides coordination for the list of standards approved for use in NASA Earth Science Data Systems and community organization through teleconferences and working groups for discussion of existing and emerging standards. Together, the DAACs provide and deploy a wide array of data discovery tools. In addition to the toolsets and capabilities already mentioned, numerous data visualization and analysis tools support a wide variety of data types and sources. Examples include Giovanni (giovanni.gsfc.nasa.gov) and MODIS subsetting and overlay tools (daac.ornl.gov/MODIS/), with an emphasis on multivariate and multitemporal remote-sensing data products. EOSDIS currently includes more than 8,000 unique data collections, with a total archive volume of 9 PB that is growing at a rate of more than 6 TB a day. The system has registered more than 2 million distinct users, with an average end-user distribution volume of 28 TB a day (statistics as of September 2014, earthdata.nasa.gov/about/system-performance/).

EOSDIS also participates in a number of national and international data community collaborations, including the Federation of Earth Science Information Partners, U.S. Group on Earth Observations, and Open Geospatial Consortium. In addition, EOSDIS actively participates in and supports the U.S. government's Climate Data Initiative (www.data.gov/climate/) and Big Earth Data Initiative.

11.2 NOAA

NOAA provides an integrated view of climate and weather data at regional to global scales through its climate.gov project, which began in 2010 as a prototyping collaboration among four NOAA offices (Climate Program Office, National Climatic Data Center, Coastal Services Center, and Climate Prediction Center). The "Maps and Data" section of climate.gov is developing to support storage, retrieval, and graphical presentation of climate and weather-related data from across NOAA and its partners' data centers. Science and data panels guide the evolution of climate.gov, with membership from within NOAA, universities, and other agencies. The data panel, which includes senior data managers from major Earth system data centers, provides input on available data sets and current and emerging technologies for data search and delivery. A relatively small number of well-curated data sets are presented with great attention to graphical formats and clear documentation, targeting a broad range of users, including scientists, policy-makers, and educators.

NOAA recently merged three major data centers (National Climatic Data Center, National Geophysical Data Center, and National Oceanographic Data Center) into a single distributed system, the National Centers for Environmental Information (NCEI). Atmospheric, oceanographic, coastal, and geophysical data products and services are being organized using a common set of data service technologies provided through common interfaces. Coverage includes data products at both national and global scales, and NCEI services target a broad user base in research and application areas. NCEI

partners with climate.gov, NOAA's National Weather Service (weather.gov), the National Integrated Drought Information System (drought.gov), and the U.S. Global Change Research Program (www.globalchange.gov).

Other parts of NOAA support additional data services, such as the National Centers for Environmental Prediction, which maintains and distributes a wide range of climate-relevant information, and the Geophysical Fluid Dynamics Laboratory, which supports search, retrieval, and distribution of climate modeling data, including implementation of an Earth System Grid Federation node.

11.3 NSF

In response to the Big Data Initiative announced by the White House Office of Science and Technology Policy in 2012, NSF has invested in multiple efforts, including the Data Observation Network for Earth (DataONE), EarthCube, and a project integrating Algorithms, Machines, and People (AMP).

DataOne (www.dataone.org) is intended to provide a single point of access to a broad range of data resources, drawing together a metacollection of Earth data from many partners. A working group structure provides guidance on current and emerging efforts connected to the lifecycle of large and complex data systems. Working groups include Sustainability and Governance, Community Engagement and Outreach, Cyberinfrastructure, and Usability and Assessment. Data search capabilities link users to one or more of the current 27 member "nodes." In addition to data access, DataONE also provides and updates detailed information on best practices for data management and maintains a compilation of useful software tools. DataONE partners with the Data Management Planning Tool (dmptool.org) to provide resources for creating, reviewing, and sharing data management plans.

EarthCube, supported by both the Geosciences and Advanced Cyberinfrastructure programs in NSF, seeks to increase the availability of data and associated tools and services in the broad Earth sciences community, increasing knowledge availability for society as a long-term goal.

AMP (amplab.cs.berkeley.edu) addresses scientific challenges related to applying newly available large-scale computing resources to the burgeoning volume of data and growing requirements for data analysis. Variable data quality, formats, and sources make applying traditional analysis algorithms to the largest data sets difficult, and available computer architectures are not always compatible with current algorithmic and programming models. Machine learning, data mining, language processing, and speech recognition are all areas being explored under AMP as avenues for improved knowledge discovery.

11.4 Opportunities for Coordination

The large data and informatics efforts summarized above are just a few of many efforts currently under way in this domain. A comprehensive list is beyond the scope of the workshop or this report but would include dozens of agencies and institutions at the local, state, national, and international levels. Beyond developing a more complete awareness of this broad landscape and a refined appreciation for the capabilities and expertise available in different agencies and centers, it is also necessary to define strategic partnerships that meet DOE Office of Biological and Environmental Research objectives while providing an added value to a broad and growing data and informatics community. Some of this coordination will take place at the level of agency and organizational representatives, but there is also a role for data management practitioners and data center operations specialists, in coordination with science team representatives across a range of projects and agencies, to develop system requirements and suggest creative adaptations and reconfigurations of existing efforts to meet those requirements. If these integration efforts can reach across agency and institutional boundaries, efficiencies of scale and leveraging of unique capabilities likely will emerge. This workshop report should be seen as one step toward the realization of that broader objective.

12. References

ASCAC. 2013. *Synergistic Challenges in Data-Intensive Science and Exascale Computing; Summary Report of the Advanced Scientific Computing Advisory Committee (ASCAC) Subcommittee*, U.S. Department of Energy Office of Science (science.energy.gov/~/media/40749FD92B58438594256267425C4AD1.ashx).

BERAC. 2013. *BER Virtual Laboratory: Innovative Framework for Biological and Environmental Grand Challenges; A Report from the Biological and Environmental Research Advisory Committee*, DOE/SC-0156. U.S. Department of Energy Office of Science (genomicscience.energy.gov/program/beracvirtuallab.shtml).

Chard, K., S. Tuecke, and I. Foster. 2014. "Efficient and Secure Transfer, Synchronization, and Sharing of Big Data." *Cloud Computing, IEEE* **1**(3), 46–55. DOI:10.1109/MCC.2014.52.

Dart, E., et al. 2013. "The Science DMZ: A Network Design Pattern for Data-Intensive Science." In *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis* (SC '13). ACM, New York, NY, Article 85, 10 pp. DOI:10.1145/2503210.2503245.

U.S. DOE. 2014. *Accelerated Climate Modeling for Energy: Project Strategy and Initial Implementation Plan*. ACME Council, Office of Biological and Environmental Research, U.S. Department of Energy Office of Science (climatemodeling.science.energy.gov/sites/default/files/publications/acme-project-strategy-plan.pdf).

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

Williams, D. N., et al. 2014. "Department of Energy Strategic Roadmap for Earth System Science Data Integration." In *Proceedings of the 2014 IEEE International Conference on Big Data*. Washington, D.C., pp. 772–77. DOI:10.1109/BigData.2014.7004304.

Williams, D. N., et al. 2015. "A Global Repository for Planet-Sized Experiments and Observations," *Bulletin of the American Meteorological Society*. DOI:10.1175/BAMS-D-15-00132.1.

Appendix 1. Workshop Findings

Workshop participants reviewed current practices and future plans for multiple U.S. Department of Energy (DOE) Climate and Environmental Sciences Division (CESD) science projects in the context of the challenges facing both the "Virtual Laboratory" and data infrastructure. Attendees drew findings from workshop presentations and reports, expert testimony, and use cases. Data-intensive activities are increasing in all CESD science endeavors, and high-performance computing (HPC) facilities are important enablers of these activities. Key findings from the attendees are briefly summarized below from the perspective of identifying investments that are most likely to positively impact CESD science goals and missions.

Topic	Finding
Interagency Partnerships	The challenges of distributed Big Data management and analysis are too large for CESD to solve alone. CESD will only succeed in its exciting Virtual Laboratory goals by leveraging best-of-breed research data management technologies used by the science community.
Reproducibility and Repeatability	The sheer size of current and expected future archives makes the storage and analysis of data on users' personal workstations impossible. Therefore, researchers need the ability to submit complex data analysis workflows that seamlessly process data stored at distributed locations. Detailed metadata and provenance information about the workflow (e.g., inputs, outputs, and algorithms) must be captured and made publicly available so that other researchers can fully understand and reproduce or repeat the results.
Resource Funding	DOE researchers have led the world in the application of advanced computing to computational simulation. In contrast, DOE climate and environmental sciences suffer from an *ad hoc*, under-resourced research data infrastructure. This situation significantly hinders progress in research programs of great scientific and societal importance. For example, availability and reliability of hardware and other resources for data analysis are major issues.
Storage and Computing	CESD currently lacks the storage and computing resources required to achieve its science goals. CESD should establish strong strategic partnerships with the DOE Office of Advanced Scientific Computing Research (ASCR) to ensure availability of those resources and examine the feasibility of using commercial cloud resources for some purposes.
Scalability	Services must be designed to scale to the order of magnitude expected of future data and metadata archives in the next 5 to 10 years, while still guaranteeing a satisfactory level of performance for users. In particular, the infrastructure must be able to support the hundreds of petabyte-sized distributed archival data that are expected to be produced by the next generation of climate models and higher-resolution observational instruments.
Proactive Engagement with CESD Projects	Projects and development teams must continuously and proactively engage with all possible areas of the project or programs: data users and providers, project coordinators, infrastructure providers, and funding program managers. This engagement will guarantee that data, software, and resources are developed and utilized to fulfill stakeholder requirements, maximize user satisfaction, and achieve the expected level of service.
Model Runs	Model developers and modelers have varied data management needs. Requirements include (1) the ability to perform many small model runs with rapid turnaround during the model development phase, (2) more computationally demanding uncertainty quantification and optimization work for model refinement, and (3) massive data runs on leading supercomputers with the full array of analysis, diagnostics, and model metrics features once the models are in production. Modelers are expected to use shareable, reproducible, or repeatable workflows; access data from many heterogeneous data sources; and run HPC *in situ* analyses, diagnostics, and model metrics.

Topic	Finding
Data Transfers	When conducting large-scale analyses of data sets from multiple climate models, the data sets typically are assembled at an HPC facility where the scientist has the necessary computing allocation to run the analysis. This process requires high-performance data transfer capabilities among major data centers and HPC facilities, which have the necessary computing and storage capabilities to support these large-scale *in situ* analyses. Therefore, it is critical that the data centers and HPC facilities support the transfer of large-scale data sets to major computing facilities in addition to data subsetting and co-located data analysis services. Researchers currently spend an enormous amount of time thinking about where the data are physically located and how to co-locate data for analysis.
Uncertainty Quantification (UQ)	A system that cross-references uncertainty estimates on observational and modeling results is needed to ensure that empirical constraints are applied appropriately. Analysis of large and multidimensional model outputs is required to interpret UQ results. Filtering of sensitivity analysis results produces a reduced set of parameters for formal estimation, but these results can vary in space and time, placing high demands on the analysis framework and requiring engagement of expert knowledge.
Data Access and Ownership	Scientific projects, data providers, and users expect a reliable data infrastructure to make their products visible and accessible, while enabling them to control and track product utilization and receive appropriate credit for their contributions. Data should be clearly identifiable and recognizable via digital object identifiers (DOIs), and owners of the data must be recognized, possibly by open researcher and contributor identifiers (ORCIDs).
Discovery	Finding physically related data among programs and projects in CESD is difficult.
Standardization	Whenever possible, the infrastructure must conform to established standards for flexible interactions. This also will maximize interoperability with data systems and facilities of other U.S. and international agencies (e.g., National Aeronautics and Space Administration, National Oceanic and Atmospheric Administration, and National Science Foundation). Additionally, interoperability greatly increases the level of user satisfaction because users are not compelled to learn and develop different techniques to access services from different systems.
Data Movement	High-speed, reliable data movement is essential to Virtual Laboratory goals. CESD should work closely with the Energy Sciences Network (ESnet) and Globus to ensure high-speed, reliable, and secure end-to-end communications among its researchers, facilities, and other relevant resources.
Best Data and Software Practices	Software and service reliability is frequently underemphasized in science but is absolutely essential to Virtual Laboratory goals. CESD must ensure that future programs leverage best-practice methods to achieve the high reliability required to meet science goals. Achieving this goal means developers must strive to apply recommended best practices in all phases of the data and software lifecycle (i.e., design, development, testing, deployment, and operation) and across all software layers (see Fig. 2, p. 17). Best practices can be achieved by many collaborative events, such as software code sprints, code reviews, and test coverage analysis, and also involve common data curation policies across CESD and various agencies.
Search and Discovery	A user must be able to search, discover, download, and analyze data hosted at different centers and facilities as if the data were served from a single location. The distributed nature of the system must be totally transparent to end users and clients, establishing common metadata for raw and post-processed data across CESD to facilitate search and discovery.
Monitoring and Metrics	To permit continuous, data-driven improvement of its operations and investments, the Virtual Laboratory should incorporate extensive monitoring and logging capabilities to permit detailed and accurate analysis of its performance, reliability, security, and usage. This environment also includes facilities for capturing and analyzing metrics about utilization of services, as well as for estimating the impact of the data infrastructure throughout the science community (e.g., as quantified by the number of science papers that use data sets downloaded from the infrastructure or based on processing algorithms executed on ASCR servers). These metrics can be used to both improve the performance and quality of services and report usage to CESD program mangers.
Modularity	The infrastructure should not be built as a monolithic package that must be installed and upgraded as a whole. Rather, it should be based on the integration of several servers and libraries meant to be upgraded and possibly replaced individually (see Fig. 2, p. 17). This philosophy enables the infrastructure to continuously evolve to incorporate new advances in all classes of services such as data discovery, transfer, analysis, and visualization.

Topic	Finding
Local or Remote and *in situ* Analysis	Server-side and *in situ* computation is necessary as increases in data size and algorithm complexity lead to data- and compute-intensive challenges for diagnostics, UQ, analysis, model metrics, and visualization. For complete flexibility, the analysis system must abstract a data file's physical location and let the back-end dynamic resource manager decide how, when, and where to move the data for small- and large-scale analyses. These requirements include the creation of a cloud-based CESD analysis platform that can scale to the needs of CESD scientists.
Unified Access Control	In particular, a user or client must not be asked to authenticate or be authorized separately at all data centers or ASCR facilities. Rather, the system infrastructure must support "single sign-on" for authentication and federated access control, whereby the authorization statements issued by one center are honored by other peer centers to access the same class of resources.
Operations	Data management, stewardship, and curation are ongoing, long-lived functions requiring a strategy that is resilient to continuing evolution in hardware and software.

Appendix 2. Workshop Example Questions

Data Infrastructure
How do we integrate all data holdings within the Climate and Environmental Sciences Division (CESD) and, eventually, the DOE Office of Biological and Environmental Research (BER)?
What are the missing components that need to be developed to integrate existing BER data archives?
What type of construct should be used (e.g., co-located or federated)?
Should this integrated environment construct be a facility or a project?
Can this construct be complementary to existing data efforts supported by other agencies (e.g., EarthCube and the National Aeronautics and Space Administration's (NASA) Distributed Active Archive Centers (DAACs)? If so, how?
How should the data be served to CESD's research communities? • What modes of data transfer should be available to the users of this system?
Should a simple compute visualization framework be incorporated? • What should its capabilities be? • Should these calculations be done locally or server side?
Compute Environment
How will BER scientists be doing research and interacting with large volumes of data 10+ years from now?
What type of data and computing environment will be necessary for this seamless integration?
Will supporting this type of system be possible within a heterogeneous compute environment?
Will task automation be a necessary component of this system?
How will code reusability be addressed within this construct?
Will exascale compute resources be a necessary component or a complementary resource for this structure? • Regardless of where this system is implemented, it must appear transparent to a user. Which necessary components must be addressed to make this happen? • Which components are key failure points?

Appendix 3. Survey Questions — Overall Ranking

The new DOE mandate on data management and sharing clearly has penetrated the research community and raises questions for many, as indicated by high survey scores for a number of related requests such as the following:

- Easy way to publish and archive data using one of the DOE data centers (received highest score overall, with a 4.79 average rating; nearly 70% of responders identified this as their highest or second-highest requirement).
- User support for data access and usage (high rating of 4.64).
- Access to enough computational and storage resources (similarly high interest, with ratings of 4.52 and 41%).

The increased interest in collaborative environments for the sharing of data and information within and between scientific groups also could be tied into this topic area.

Survey Question	Average Rating
Easy way to publish and archive data using one of the DOE data centers	4.79
Means for comparing diverse data types generated from observations and simulations	4.71
User support for data access and usage	4.64
Access to sufficient observational and experimental resources	4.58
Access to enough computational and storage resources	4.52
Ingestion and access to large volumes of scientific data (i.e., from a data archive to supercomputer)	4.49
Quality control algorithms for data	4.46
A unified and single user account to access all DOE resources within BER and the Office of Advanced Scientific Computing Research (ASCR)	4.44
In situ analysis of observational, experimental, and computational results. Ability to interpret results and verify new insights within the context of existing scientific knowledge	4.40
Means of comparing data collected at different scales	4.34
Collaborative environments	4.31
Availability of ancillary data products such as data plots, statistical summaries, data quality information, and other documentation	4.22
Rapid data quality assessment during discovery	4.18
Interoperability: Interfaces that ensure a high degree of interoperability for different formats and semantic levels among repositories and applications	4.18
Data manipulation before download (e.g., averaging and subsetting)	4.16
Capture of provenance information for data	4.11
Reproducibility	4.06
Libraries and repositories that allow for community-wide authentication and access across institutions and communities	4.06
Improved user interfaces	4.00
Unified data discovery for all BER data sources to support user research	4.00
Software that enables small teams to engage in large-scale ensemble and uncertainty quantification simulations	3.88
Direct data delivery into ASCR computing systems from BER data resources	3.86
Software to ensure workflow resilience and recovery from errors	3.85
Data visualization tools	3.85

Survey Question	Average Rating
Real-time data quality control during data collection	3.74
Support for the creation of scientific workflows	3.68
Data intention: Methods and languages for describing and adhering to intellectual property in systems where not all the data are openly available	3.57
Real-time access to live data streams	3.25
New techniques to work with deep-memory hierarchies on extreme-scale computing systems	3.24

Appendix 4. Workshop Agenda

Time	Topic
Thursday, August 13, 2015	
8:45 a.m. – 9:10 a.m.	**Welcome and Introduction** (Gary Geernaert) **Workshop Charge** (Jay Hnilo)
9:10 a.m. – 9:30 a.m.	**Survey Responses** (Kerstin Kleese van Dam)
9:30 a.m. – 9:45 a.m.	**Identifying CESD Computational and Data Environment** (Dean N. Williams, Giriprakash Palanisamy)
9:45 a.m. – 10.00 a.m.	**Break**
10:00 a.m. – 11.00 a.m.	**Science Drivers** Discussion Lead (Peter Thornton) • Example use case requirements (Jay Hnilo), 10 min. • Define the key things that are difficult to do today and are impeding scientific progress or productivity • Science case discussion, 50 min. List science drivers. Assignment: convert science drivers to use cases
Team Member Lists	
Red Team Members	• David C. Bader (LLNL), Modeler • Forrest M. Hoffman (ORNL), Modeler • Deb Agarwal (LBNL), Data Management • Robert Jacob (ANL), Data Scientist • Timothy Scheibe (PNNL), Data Management • Margaret Torn (LBNL), Data Scientist • Andrew Vogelmann (BNL), Modeler • David Skinner (LBNL), Data Center • Scott M. Collis (ANL), Data Scientist
Blue Team Members	• Philip J. Rasch (PNNL), Modeler • Paul J. Durack (LLNL), Data Scientist • Peter Thornton (ORNL), Modeler • Michael F. Wehner (LBNL), Data Scientist • Thomas A. Boden (ORNL), Data Management • Jennifer Comstock (PNNL), Data Scientist • Shaocheng Xie (LLNL), Data Scientist • Mallikarjun Shankar (ORNL), Data Center • Eli Dart (ESnet), Data Center
Breakout Sessions	
11:00 a.m. – 12:30 p.m.	**Data Services to Support Science Requirements** **Red Team:** Discussion Lead (Forrest Hoffman) **Blue Team:** Discussion Lead (Shaocheng Xie) **Questions:** • What are the key challenges that scientists encounter? • Which data services would address the identified challenges? What exists already today? What is still needed? What are the key characteristics that these services need to have to be successful (e.g., integrated and easy to customize)? • What are the key impediments (on data provider or service provider side) in delivering these services? • Which services should be developed with the highest priority, and what would be their measurable impact on science?

Time	Topic
12:30 p.m. – 1:30 p.m.	**Lunch**
1:30 p.m. – 2:30 p.m.	**Breakout Session Reports and Discussion**, 30 min. per team
2:30 p.m. – 4:00 p.m.	**Required Data Center and Interoperable Services** **Red Team:** Discussion Lead (Margaret Torn) **Blue Team:** Discussion Lead (Thomas Boden) **Discuss top-priority services required to meet the community's needs as part of an integrated infrastructure, including topics such as:** • Data integration and advanced metadata capabilities • Data and metadata collection and sharing capabilities • Data quality, uncertainty quantification, and ancillary information • Use of broader ontology for discovery and use of CESD data sets • Data discovery and access, data downloading, and subsetting services and capabilities • Data preparation services and tools • Authentication and security • Local and remote publication services • Local and remote catalog and search services, data transfer services • Human-computer interface (e.g., user interface, application program interfaces) • Resource discovery and allocation services • Workflow services (link together scientific or project execution) • Computing services • Exploration services (includes analytics and visualization) • Identification and prioritization of key gaps, benefitted communities
4:00 p.m. – 4:15 p.m.	**Break**
4:15 p.m. – 5:15 p.m.	**Breakout Session Reports and Discussion**, 30 min. per team
Friday, August 14, 2015	
8:30 a.m. – 10:00 a.m.	**Advanced Computational Environments and Data Analytics** **Red Team:** Discussion Lead (Scott Collis) **Blue Team:** Discussion Lead (Paul Durack) **Questions:** • What are the key challenges that scientists encounter? • What capabilities would address the identified challenges? • What exists already today? • What do we still need? • What are the impediments for resource providers and software developers to provide these missing capabilities? • Which requirements need to be addressed with the highest priority, and what would be their measurable impact on science? **Possible discussion topics:** • Definition of a scalable compute resource (clusters and high-performance computers) for CESD data analysis • Data analytical and visualization capabilities and services • Analysis services when multiple data sets are not co-located • Performance of model execution • Advanced networks as easy-to-use community resources • Provenance and workflow • Automation of steps for the computational work environment • Resource management, installation, and customer support • Identification and prioritization of key gaps and benefitted communities
10:00 a.m. – 10:15 a.m.	**Break**
10:15 a.m. – 10:45 a.m.	**Breakout Session Reports and Discussion,** 15 min. per team

Time	Topic
10:45 a.m. – 11:45 a.m.	**Inventory of existing CESD data tools and services, benchmark of tools for potential reuse** **Red Team:** Discussion Lead (Deb Agarwal) **Blue Team:** Discussion Lead (Jennifer Comstock) **Suggested subtopics:** • What tools have been identified during the previous discussions that should be made more widely accessible to the CESD community? • What other existing tools could address key needs? • How should tools and services be made available today and in the future in an integrated infrastructure? What level of support would be expected from the science community? • How should tool maturity and capabilities be assessed (e.g., benchmarks or crowd sourcing)? • Are there any conventions needed for your project?
11:45 a.m. – 12:15 p.m.	**Breakout Session Reports and Discussion**, 15 min. per team
12:15 p.m. – 1:15 p.m.	**Lunch**
1:15 p.m. – 2:00 p.m.	**General Discussion: Data Services and Monitoring** Discussion Lead (Eli Dart) **Questions:** • What level of service, monitoring, maintenance, and metrics are needed for data services and tools? • What do service providers want to see from others? • To what do the scientists want access?
2:00 p.m. – 2:30 p.m.	**General Discussion: Participation with Broad, Multiagency Data Initiatives** Discussion Lead (Peter Thornton) **Suggested subtopics:** • Standards and services that need to be adopted within the compute environment that will allow CESD to participate in multiagency data initiatives such as EarthCube and the U.S. Group on Earth Observations • Data sharing with NASA's DAACs, the National Oceanic and Atmospheric Administration, and other agencies
2:30 p.m. – 3:00 p.m.	**Summary of Action Items, Workshop Report Draft** **Follow-Up and Future Workshop Ideas**

Appendix 5. Workshop Participants

Name	Area of Representation	Affiliation	Email Address
Participants			
Agarwal, Deb	Earth and Environmental Sciences	LBNL	daagarwal@lbl.gov
Bader, David C.	ACME, LLNL Climate Science	LLNL	bader2@llnl.gov
Boden, Thomas A.	AmeriFlux, CDIAC, FACE, NGEE	ORNL	bodenta@ornl.gov
Collis, Scott M.	HPC, Py-ART, Radar	ANL	scollis@anl.gov
Comstock, Jennifer	ARM, ASR	PNNL	jennifer.comstock@pnnl.gov
Dart, Eli	ESnet	ESnet	dart@es.net
Durack, Paul J.	PCMDI, MIPs, RGCM	LLNL	durack1@llnl.gov
Hoffman, Forrest M.	ILAMB, ACME	ORNL	hoffmanfm@ornl.gov
Jacob, Robert	HPC, ACME	ANL	jacob@mcs.anl.gov
Kleese van Dam, Kerstin *	EMSL, ARM	PNNL/BNL	kerstin.kleesevandam@pnnl.gov
Palanisamy, Giriprakash *	ARM, NGEE	ORNL	palanisamyg@ornl.gov
Rasch, Philip J.	ACME	PNNL	philip.rasch@pnnl.gov
Scheibe, Timothy	EMSL	PNNL	tim.scheibe@pnnl.gov
Shankar, Mallikarjun	OLCF	ORNL	shankarm@ornl.gov
Skinner, David	NERSC	LBNL	deskinner@lbl.gov
Thornton, Peter	ACME, NGEE	ORNL	thorntonpe@ornl.gov
Torn, Margaret S.	AmeriFlux, ASR	LBNL	mstorn@lbl.gov
Vogelmann, Andrew	ARM	BNL	vogelmann@bnl.gov
Wehner, Michael F.	CASCADE	LBNL	mfwehner@lbl.gov
Williams, Dean N. *	ACME, MIPs, ESGF	LLNL	williams13@llnl.gov
Xie, Shaocheng	ACME, ARM, RGCM/ASR (CAPT)	LLNL	xie2@llnl.gov
Participants from DOE Program Offices			
Bayer, Paul	Program Manager	BER	paul.bayer@science.doe.gov
Geernaert, Gary	BER CESD Director	BER	gary.geernaert@science.doe.gov
Hnilo, Justin *	Program Manager	BER	justin.hnilo@science.doe.gov
Joseph, Renu	Program Manager	BER	renu.joseph@science.doe.gov
McFarlane, Sally	Program Manager	BER	sally.mcfarlane@science.doe.gov
Ndousse-Fetter, Thomas	Program Manager	ASCR	thomas.ndousse-fetter@ science.doe.gov
Petty, Rickey	Program Manager	BER	rick.petty@science.doe.gov

** Workshop and report co-chairs and organizers*

Appendix 6. Acronyms, Abbreviations, and Terms

Acronym	Description
ACME	Accelerated Climate Modeling for Energy — DOE's effort to build an Earth system modeling capability tailored to meet climate change research strategic objectives (climatemodeling.science.energy.gov/projects/accelerated-climate-modeling-energy).
ALCF	Argonne Leadership Computing Facility —DOE Office of Science user facility that provides researchers from national laboratories, academia, and industry with access to high-performance computing capabilities (www.alcf.anl.gov).
AmeriFlux	AmeriFlux — Community of sites and scientists who measure ecosystem carbon, water, and energy fluxes across the Americas and are committed to producing and sharing high-quality eddy covariance data (AmeriFlux Site and Data Exploration System, ameriflux.ornl.gov).
AMP	NSF Algorithms, Machines, and People project — Works at the intersection of machine learning, cloud computing, and crowdsourcing to create a new Big Data analytics platform (amplab.cs.berkeley.edu).
ANL	Argonne National Laboratory — Science and engineering research national laboratory near Lemont, Illinios, operated by the University of Chicago for DOE (www.anl.gov).
API	application program interface (en.wikipedia.org/wiki/Application_programming_interface/).
ARM	Atmospheric Radiation Measurement — The ARM Climate Research Facility is a DOE user facility that provides *in situ* and remote-sensing observations to improve the understanding and climate model representations of clouds, aerosols, and their interactions with Earth's surface (www.arm.gov).
ASCR	DOE Office of Advanced Scientific Computing Research — Discovers, develops, and deploys computational and networking capabilities to analyze, model, simulate, and predict complex phenomena important to DOE (science.energy.gov/ascr/).
ASR	Atmospheric System Research — DOE BER program that advances process-level understanding of the key interactions among aerosols, clouds, precipitation, radiation, dynamics, and thermodynamics to reduce the uncertainty in global and regional climate simulations and projections (science.energy.gov/ber/research/cesd/atmospheric-system-research-program/).
BER	DOE Office of Biological and Environmental Research — Supports world-class biological and environmental research programs and scientific user facilities to faciliate DOE's energy, environment, and basic research missions (science.energy.gov/ber/).
BERAC	Biological and Environmental Research Advisory Committee (science.energy.gov/ber/berac/).
BNL	Brookhaven National Laboratory — National research institution in Upton, New York, funded primarily by DOE that provides expertise and world-class facilities for studies in physics, chemistry, biology, medicine, applied science, and a wide range of advanced technologies (www.bnl.gov).
CADES	Compute and Data Environment for Science — Oak Ridge National Laboratory compute and data infrastructure coupled with data science experts focused on creating a data-centric environment for scientific discovery.
CASCADE	Calibrated and Systematic Characterization, Attribution, and Detection of Extremes (cascade.lbl.gov)
CDIAC	Carbon Dioxide Information Analysis Center — DOE's primary climate change data and information analysis center whose data holdings include estimates of CO_2 emissions from fossil fuel consumption and land-use changes, records of atmospheric concentrations of CO_2 and other radiatively active trace gases, carbon cycle and terrestrial carbon management data sets and analyses, and global and regional climate data and time series (cdiac.ornl.gov).
CDO	climate data operators — Collection of command-line operators to manipulate and analyze climate and numerical weather prediction data.
CESD	Climate and Environmental Sciences Division — Within DOE's Office of Biological and Environmental Research, CESD focuses on advancing a predictive understanding of Earth's climate and environmental systems to inform the development of sustainable solutions to U.S. energy and environmental challenges (science.energy.gov/ber/research/cesd/).
CF	Climate forecast conventions and metadata (cfconventions.org).
CH_4	methane

Acronym	Description
CMIP	Coupled Model Intercomparison Project — Sponsored by the World Climate Research Programme's Working Group on Coupled Modeling, CMIP is a community-based infrastructure for climate model diagnosis, validation, intercomparison, documentation, and data access (cmip-pcmdi.llnl.gov).
CMR	Common Metadata Repository — An earth science metadata repository for NASA EOSDIS data (earthdata.nasa.gov/about/science-system-description/eosdis-components/common-metadata-repository).
CO_2	carbon dioxide
DAACs	NASA Distributed Active Archive Centers — Process, archive, document, and distribute data from NASA's past and current Earth-observing satellites and field measurement programs (earthdata.nasa.gov/about/daacs).
DataONE	NSF Data Observation Network for Earth — Distributed framework and sustainable cyberinfrastructure for open, persistent, robust, and secure access to well-described and easily discoverable Earth observational data (www.dataone.org).
DOE	U.S. Department of Energy — Government agency chiefly responsible for implementing energy policy (www.energy.gov).
DOI	digital object identifier — A serial code used to uniquely identify content of various types on electronic networks; particularly used for electronic documents such as journal articles (en.wikipedia.org/wiki/Digital_object_identifier/).
DTN	data transfer node — Internet location providing data access, processing, or transfer.
EMSL	DOE Environmental Molecular Sciences Laboratory — National scientific user facility providing scientific expertise, instruments, and capabilities in support of research to predictively understand the molecular-to-mesoscale processes in biological, climate, environmental, and energy systems (www.emsl.pnl.gov/emslweb/).
EOS	NASA Earth Observing System — Coordinated series of polar-orbiting and low inclination satellites for long-term global observations of the land surface, biosphere, solid Earth, atmosphere, and oceans (eospso.nasa.gov).
EOSDIS	NASA Earth Observing System Data and Information System — Provides end-to-end capabilities for managing NASA's Earth science data from different sources, including satellites, aircraft, field measurements, and various programs (earthdata.nasa.gov).
ESDIS	NASA Earth Science Data and Information System — Manages the science systems of EOSDIS (earthdata.nasa.gov/about/esdis-project/).
ESGF	Earth System Grid Federation — Led by Lawerence Livermore National Laboratory, a worldwide federation of climate and computer scientists deploying a distributed multipetabyte archive for climate science (esgf.llnl.gov).
ESM	Earth system model — Type of complex, global model that combines physical climate models, global biological processes, and human activities.
ESnet	DOE Energy Sciences Network — Provides high-bandwidth connections that link scientists at national laboratories, universities, and other research institutions, enabling them to collaborate on scientific challenges including energy, climate science, and the origins of the universe (www.es.net).
FACE	Free-Air CO_2 Enrichment — Project designed to permit the experimental exposure of tall vegetation, such as stands of forest trees, to elevated atmospheric carbon dioxide concentrations without enclosures that alter the tree microenvironment.
FDR IB	Fourteen Data Rate InfiniBand
GB	gigabyte
Globus	Provides high-performance, secure, and reliable data transfer, sharing, synchronization, and publication services for the science community (www.globus.org).
GCMD	Global Change Master Directory — A resource for the discovery, access, and use of Earth science data and data-related services worldwide, while specifically promoting the discovery and use of NASA data. (gcmd.nasa.gov).
GridFTP	High-performance, secure, reliable data transfer protocol optimized for high-bandwidth wide-area networks (toolkit.globus.org/toolkit/docs/latest-stable/gridftp/).
HCI	human-computer interaction

Acronym	Description
HDF5	Hierarchical Data Format version 5 — Data model, library, and file format for storing and managing a wide variety of high volume and complex data types (www.hdfgroup.org/HDF5/).
HPC	high-performance computing
ICNWG	International Climate Network Working Group — Formed under the Earth System Grid Federation to help set up and optimize network infrastructure for climate data sites around the world (icnwg.llnl.gov).
IDL	Interactive Data Language — High-level programming language for data manipulation, visualization, and analysis having strong signal and image processing capabilities and extensive math and statistical functions.
ILAMB	International Land Model Benchmarking — Model-data intercomparison and integration project designed to improve the performance of land models and enhance the design of new measurement campaigns to reduce uncertainties associated with key land surface processes (www.ilamb.org).
INCITE	DOE Innovative and Novel Computational Impact on Theory and Experiment program — Accelerates scientific discoveries and technological innovations by awarding, on a competitive basis, time on supercomputers to researchers with large-scale, computationally intensive projects that address "grand challenges" in science and engineering (www.doeleadershipcomputing.org/incite-program/).
I/O	input/output
ISO	International Organization for Standardization — Independent, nongovernmental organization that brings together international experts to develop voluntary, consensus-based, market relevant international standards that support innovation and provide solutions to global challenges (www.iso.org/iso/).
KD	knowledge gathering, managing, and sharing
LANCE	Land, Atmosphere Near real-time Capability for EOS — A group of five near-real time data systems serving the land and atmosphere science community (earthdata.nasa.gov/earth-observation-data/near-real-time).
LBNL	Lawrence Berkeley National Laboratory — DOE Office of Science laboratory managed by the University of California that conducts fundamental science for transformational solutions to energy and environment challenges using interdisciplinary teams and advanced new tools for scientific discovery (www.lbl.gov).
LLNL	Lawrence Livermore National Laboratory — DOE laboratory that develops and applies world-class science and technology to enhance the nation's defense and address scientific issues of national importance (www.llnl.gov).
Metadata	Data properties, such as origin, spatiotemporal extent, and format (en.wikipedia.org/wiki/Metadata).
MIPs	model intercomparisons
NASA	National Aeronautics and Space Administration — U.S. government agency responsible for the civilian space program as well as aeronautics and aerospace research (www.nasa.gov).
NCAR	National Center for Atmospheric Research — Federally funded research and development center devoted to service, research, and education in atmospheric and related sciences (ncar.ucar.edu).
NCEI	National Centers for Environmental Information — Hosts and provides access to comprehensive archives of oceanic, atmospheric, and geophysical data (www.ncei.noaa.gov).
NCL	NCAR Command Language — Interpreted language designed specifically for scientific data processing and visualization (www.ncl.ucar.edu).
NCO	NetCDF Operators
NERSC	National Energy Research Scientific Computing Center — Primary scientific computing facility for the DOE Office of Science, providing computational resources and expertise for basic scientific research (www.nersc.gov).
NetCDF	Network Common Data Form — A machine-independent, self-describing binary data format (www.unidata.ucar.edu/software/netcdf/).
NGEE	Next-Generation Ecosystem Experiments — DOE BER concept for coupling models with experimental and observational campaigns in long-term studies examining the response of Arctic terrestrial ecosystems and tropical forest ecosystems to climate change (NGEE-Arctic, ngee-arctic.ornl.gov; NGEE-Tropics, esd.lbl.gov/ngee-tropics/).
NOAA	National Oceanic and Atmospheric Administration — Federal agency whose missions include understanding and predicting changes in climate, weather, oceans, and coasts and conserving and managing coastal and marine ecosystems and resources (www.noaa.gov).

Acronym	Description
NSF	National Science Foundation — Federal agency that supports fundamental research and education in all the nonmedical fields of science and engineering (www.nsf.gov).
OLCF	Oak Ridge Leadership Computing Facility — DOE national user facility providing the open scientific community support and access to computing resources including the nation's most powerful supercomputer to address grand challenges in climate, materials, nuclear science, and a wide range of other disciplines (www.olcf.ornl.gov).
ORCID	Open Research and Contributor Identifier — A nonproprietary alphanumeric code to uniquely identify scientific and other academic authors (en.wikipedia.org/wiki/ORCID).
ORNL	Oak Ridge National Laboratory — DOE science and energy laboratory conducting basic and applied research to deliver transformative solutions to compelling problems in energy and security (www.ornl.gov).
PB	petabyte
PCMDI	Program for Climate Model Diagnosis and Intercomparison — Develops improved methods and tools for the diagnosis and intercomparison of general circulation models that simulate the global climate (www-pcmdi.llnl.gov).
perfSONAR	PERFormance focused Service Oriented Network monitoring Architecture — Test and measurement infrastructure used by science networks and facilities around the world to monitor and ensure network performance (www.perfsonar.net).
PI	principal investigator
PNNL	Pacific Northwest National Laboratory — DOE national laboratory in Richland, Wash., where multidisciplinary scientific teams address problems in four areas: science, energy, the Earth, and national security (www.pnnl.gov).
PSUADE	Problem Solving environment for Uncertainty Analysis and Design Exploration — Software toolkit for performing uncertainty analysis, global sensitivity analysis, design optimization, and calibration of computational models (computation.llnl.gov/casc/uncertainty_quantification/).
Py-ART	Python ARM Radar Toolkit — Python module containing a collection of weather radar algorithms and utilities (arm-doe.github.io/pyart/).
QA	quality assurance
QC	quality control
RGCM	Regional and Global Climate Modeling — DOE BER program that supports research analyzing the dominant governing processes that describe regional-scale climate change; evaluating methods to obtain higher spatial resolution for projections of climate and Earth system change; and diagnosing model systems that are cause for uncertainty in regional climate projections (science.energy.gov/ber/research/cesd/regional-and-global-modeling/).
SaaS	software as a service
SSD	solid-state disk
TB	terabyte
TECA	Toolkit for Extreme Climate Analysis — Software developed at Lawrence Berkeley National Laboratory to help climate researchers detect extreme weather events in large data sets.
TES	Terrestrial Ecosystem Science — DOE BER program seeking to improve the representation of terrestrial ecosystem processes in Earth system models, thereby advancing the quality of climate model projections and providing the scientific foundation for solutions to pressing energy and environmental challenges (tes.science.energy.gov).
THREDDS	Thematic Real-time Environmental Distributed Data Services — Web server that provides metadata and data access for scientific data sets using a variety of remote data access protocols (www.dataone.org/software-tools/thematic-realtime-environmental-distributed-data-services-thredds/).
UQ	Uncertainty quantification — Method determining how likely a particular outcome is, given the inherent uncertainties or unknowns in a system (en.wikipedia.org/wiki/Uncertainty_quantification).
UV-CDAT	Ultrascale Visualization–Climate Data Analysis Tools — Provides access to large-scale data analysis and visualization tools for the climate modeling and observational communities (uvcdat.llnl.gov).
Web portal	A point of access to information on the World Wide Web (en.wikipedia.org/wiki/Web_portal/).