LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-TR-653745

Department of Energy's Biological and Environmental Research Strategic Data Roadmap for Earth System Science

D. N. Williams, G. Palanisamy, G. Shipman, T. A. Boden, J. W. Voyles

April 28, 2014

Disclaimer

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

Department of Energy's

Biological and Environmental Research

Strategic Data Roadmap for Earth System Science

Prepared for Justin Hnilo
Data and Informatics Program Manager
April 25, 2014

Authors:
Dean N. Williams (LLNL)
Giri Palanisamy, Galen Shipman, Thomas A. Boden (ORNL)
Jimmy W Voyles (PNNL)

Table of Contents

1 Objective

Rapid advances in experimental, sensor, and computational technologies and techniques are driving exponential growth in the volume, acquisition rate, variety, and complexity of scientific data. This wealth of scientifically meaningful data has tremendous potential to lead to scientific discovery. However, to achieve scientific breakthroughs, these data must be exploitable—they must be analyzed effectively and efficiently and the results shared and communicated easily within the wider Department of Energy's (DOE's) Biological and Environmental Research (BER) Climate and Environmental Sciences Division (CESD) community. The explosion in data complexity and scale makes these tasks exceedingly difficult to achieve, particularly given that an increasing number of disciplines are working across techniques, integrating simulation and experimental or observational results (see **Table 5 in Appendix 2**). Consequently, we need new approaches to data management, analysis, and visualization that provide research teams with easy-to-use and scalable end-to-end solutions. These solutions must facilitate (and where feasible, automate and capture) every stage in the data lifecycle (shown in **Figure 1**), from collection to management, annotation, sharing, discovery, analysis, and visualization. In addition, the core functionalities are the same across climate science communities, but they require customization to adapt to specific needs and fit into research and analysis workflows. To this end, the mission of CESD's Data and Informatics Program is to integrate all existing and future distributed CESD data holdings into a seamless and unified environment for the acceleration of Earth system science.

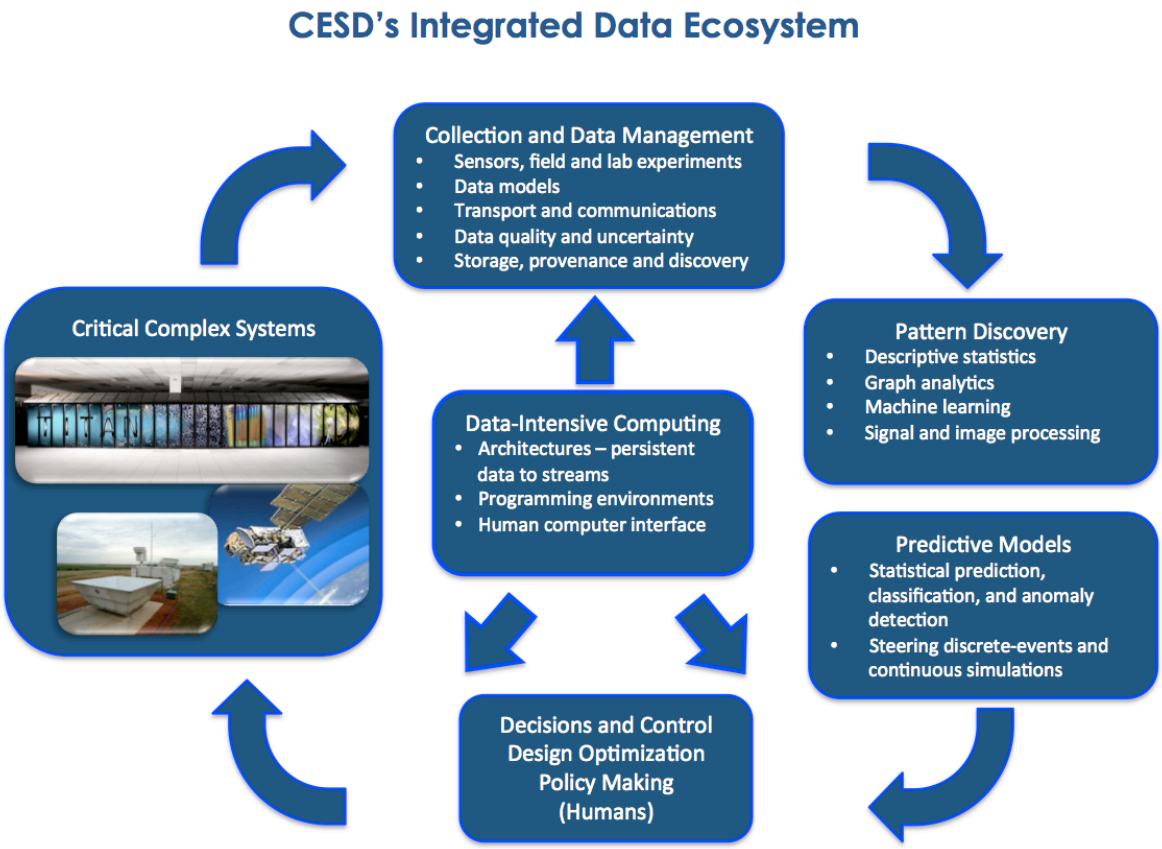

Figure 1. *The diagram depicts data flow of heterogeneous data sources as it moves through CESD's proposed integrated data ecosystem.*

This virtual laboratory ecosystem will allow science users to discover, access, and use existing heterogeneous CESD data and services with open source industry standards, protocols, and state-of-the-art technology to advance scientific discovery across disciplines in earth system science. For any existing or new BER data projects, this ecosystem will enhance the data product lifecycle through product exploration and access, research facilitation, and feedback.

Management of modeling data, because of the large scale, is pressing us to reason about storing data sets versus reproducing them. However, this reasoning does not apply to observational data, which is also increasing in volume and complexity and has the requirement to be archived in long-term repositories.

To explore complex scientific data domains and synthesize integrated research, a BER Virtual Laboratory (VL) infrastructure (see **Figure 2 in section 5**) would facilitate the product development of model initialization and comparison data products and provide a framework for the generation and reproduction of model data products. These services would be available for anyone working with or across CESD scientific data archives. The BER VL will provide automation and a record of process and analysis steps to generate repeatable beginning, intermediate, and end results.

The approach will include capturing each step in the process from configuration of the model, preparing initialization and observational intercomparison data products, and running the model. Model output and execution diagnostics will be captured and an automated analysis of results will be summarized. The resulting workflow will be recorded and automatically linked to associate workflows running at Leadership Computing Facilities (LCFs) and NERSC where the end-user can construct complex workflows to monitor model performance. Workflows and provenance will be transparent and provide optimized access to the data and the analyses. This will help manage the data transmission load on the network infrastructure.

The broad community of BER researchers will be able to efficiently access the most popular data products and get details about their applicability and utility and how they were produced. The intellectual merit of the BER Virtual Laboratory lies in the standards based architecture, integration, and intelligent network that fuse extensibility with the support of data product synthesis, data analysis, and visualization. The diverse BER community and their requirement for integrated research and scientific discovery will benefit from the development and integration of this core set of infrastructure components and services.

2 Background and Motivation

As stated by the Earth science reports, one of Earth science's most difficult challenges is managing and understanding massive amounts of global atmospheric, land, ocean, and sea-ice model data generated by ever more complex computer simulations and driven by ever larger qualitative and quantitative observations [2]. Because of rapid increases in technology, storage capacity, and networks and the need to share information, research communities are providing scientists, students, and policymakers access to federated open-source collaborative systems that everyone can use to explore, study, and manipulate large-scale data.

Many BER CESD-funded projects handle large and diverse data collections. The following are some of the key BER CESD data centers and portals. These centers are successfully managing their data using next-generation data management tools and subject matter experts across inter-laboratory and inter-

agency groups. Data management capabilities and resources from these data centers will be effectively used in building the virtual laboratory.

DOE's BER community-driven Earth System Grid Federation (ESGF) [3] was critical to the successful archiving, delivery, and analysis of the Coupled Model Intercomparison Project (CMIP), phase 3 (CMIP3) data for the International Panel on Climate Change (IPCC) Fourth Assessment Report (AR4). It was equally important in meeting the data management needs of the subsequent CMIP, phase 5 (CMIP5), which produced petascale data used for the 2013 IPCC Fifth Assessment Report (AR5), released in September 2013. Although the ESGF has been indisputably important to CMIP, its current and future impact on climate is not limited only to this high-profile project. ESGF has been used to host data for a number of other projects, over 40 so far (see **Table 4 in Appendix 2**). These data archives have been augmented with some related observational data sets from Carbon Dioxide Information Analysis Center (CDIAC) [4], atmospheric radiation measurements (ARM), and NASA satellite observations. The ESGF enterprise system is a worldwide collaboration that develops, deploys, and maintains software infrastructure for the management, dissemination, and analysis of model output and related observational data. The core management capabilities of ESGF include a software stack to publish data, search services, federated security, and large-scale data transfer interconnected via international network organizations [5].

Designated a national user facility in 2003, currently considered as the world's premier ground-based observations facility, the Atmospheric Radiation Measurement (ARM) Climate Research Facility [6] provides the climate research community with strategically located in situ and remote sensing observatories designed to improve the understanding and representation, in climate and Earth system models, of clouds and aerosols as well as their interactions and coupling with the Earth's surface. The scale and quality of the ARM Facility's approach to climate research has resulted in ARM setting the standard for ground-based climate research observations. The ARM Data Center now serves over 4,000 data products along with data quality information. These include observational data, PI data products, and value-added products.

To help users quickly find the atmospheric and climate measurements they need, the ARM data archive provides a completely new and dynamic tool for accessing and ordering data. The new ARM data discovery browser includes helpful features such as filtered search logic, multi-pass data selection, filtering data based on data quality, graphical views of data quality and availability, direct access to data quality reports, and the ever-popular data plots. In addition to the discovery interface, ARM data center provides a wealth of data management tools and services, such as the OME, ARM data integration tool, data quality assessment and distribution, data monitoring tools, digital object identifiers, ARM Radar data processing and visualization clusters and interactive Web data visualization (NCVWeb) [7].

CDIAC is the DOE's primary climate-change data and information analysis center. CDIAC provides scientific and data management support for projects sponsored by a number of agencies, including the AmeriFlux Network; continuous observations of ecosystem level exchanges of CO_2, water, energy, and momentum at different time scales for sites in the Americas; the Ocean CO_2 Data Program of CO_2 measurements taken aboard ocean research vessels; DOE-supported FACE experiments, which evaluate plant and ecosystem response to elevated CO_2 concentrations; and the HIPPO project, which is analyzing the atmospheric carbon cycle and greenhouse gas concentrations from pole to pole over the Pacific Ocean.

Over the past several years, there have been numerous workshops and reports that have highlighted the increasing size and complexity of scientific data produced by modern science in general [8, 9], DOE

facilities [10, 11], and the Earth science community, in particular BER [12, 13]. These efforts illustrate use cases and state the need for reliable, scalable collaborative infrastructures to enable effective analysis of data to further scientific discovery [14, 15]. In addition, BER conducted a one-day data workshop on June 26, 2012, that resulted in two uses cases:

- Use Case I: PhD student Miss New-user is developing an Earth System Model of intermediate complexity for a particular application and needs to test the results of her model against available CMIP and observational data. What data and model results are available and where can she find them?
- Use Case II: Program scientist Dr. NGEE Field Scientist is generating data sets and model outputs and needs to perform integration and analysis, potentially with access to other BER data resources in other programs. What platform and tools could be used to find these resources and perform the analytical functions?

To address use case I, a multi-lab team developed the BER CESD Data Gateway prototype where users can discover and access any BER CESD funded and related data sets. The initial prototype included data sets from ESGF, ARM and CDIAC. The gateway provides metadata publishing capabilities for BER CESD data projects and various data search capabilities for end users. This tool also provides seamless access to visualization, subsetting, and data-download tools presently served by the participating data centers and projects (see **Tables 1 and 2 in Appendix 2)**. It also offers metadata quality assessment and feature keyword enhancements using semantic mappings.

Although not given as examples, this white paper also addresses the architecture needs for use case II and assist in the developing components of BER's VL infrastructure.

3 Scientific Data Landscape

Earth science is an example of a discipline in which scientific progress is critically dependent on the availability of a reliable infrastructure for managing and accessing often large and heterogeneous quantities of data on a global scale. Advancing Earth science is inherently a collaborative and multi-disciplinary effort that requires sophisticated modeling of the physical processes and exchange mechanisms among multiple Earth realms (atmosphere, land, ocean, and sea ice) and comparison and validation of these simulations with observational data from various sources, possibly collected over long periods of time.

For the past decades, the climate community has worked on concerted, worldwide modeling activities led by the WGCM, sponsored by the WCRP, which led to successive reports by the IPCC. These activities involve tens of modeling groups in as many countries, running the same prescribed set of climate change scenarios on the most advanced supercomputers and producing several petabytes (PB = 10^{15} bytes) of output containing hundreds of physical variables spanning tens and hundreds of years. These data sets are held at distributed locations around the globe but must be discovered, downloaded, and analyzed as if they were stored in a single archive, with efficient and reliable access mechanisms that can span political and institutional boundaries.

Similarly, observational facilities such as ARM sites, AmeriFlux sites and Earth observing satellites are successfully collecting and disseminating very-complex and diverse observational data to improve the scientific understanding of global climate change and promote the advancement of climate models.

In addition, DOE BER CESD for decades has supported intensive field campaigns and experiments to

better understand ecosystem and biogeochemical processes. These experimental and field campaign data are needed to produce better model parameters and initializations, and to test models. The need for providing data products on demand, as well as value-added products, adds another dimension to the needed capabilities.

Finally, science results must be applied at multiple scales (global, regional, and local) and made available to different communities (scientists, policy makers, instructors, farmers, and industry). Because of its high visibility and direct impact on political decisions that govern human activities, the end-to-end scientific investigation must be completely transparent, collaborative, and reproducible. Scientists must be given the environment and tools to work with colleagues in opposing time zones, including exchanging ideas, investigating metadata, tracking provenance, annotating results, and collaborating in developing analysis applications and algorithms. The BER VL is essential for advancing scientific discovery in Earth science.

Table 2 in Appendix 2 lists the data products, services, gateways and portals, and projects that assist the efforts of the BER CESD community to better understand the sciences related to the planet Earth. The BER VL will bring this diverse group of assets together under a single project.

4 Data and Informatics Program Primary Goals

The primary goal for CESD's data infrastructure and BER's VL is to standardize and use the most up to date tools to integrate CESD's diverse data holdings and to provide data and information technology resources focused on delivering optimal scientific data and models. CESD's data management component provides the scientific interface through which technical information can be obtained, evaluated, quality-assured, documented, and distributed; the exchange of data can be promoted and facilitated; and high-quality analyses of complex data can be performed to synthesize information used in evaluating environmental issues. Specific objectives of CESD's information and integration component include promoting networking among members of the scientific community, preparing technical and informational reports, and sponsoring scientific conferences.

As an example use case, the architecture will develop a simplified workflow for candidate users. First, a user issues a scientific question on specific properties of data contained in any (or all) of the data centers. The system will then distribute partial queries to the participating individual data centers. These data centers will utilize their own data mining resources to compute the partial result. The BER CESD data analytical framework then fetches partial raw data results from each data center and translates each result into a common, understandable data format. The results are then augmented by specific domain rules input by domain experts and fused together over common attributes. The resulting data product is potentially an aggregation of the partial results that formulate the answer to the scientific question. The result is presented to the user as a reusable data product.

5 Data Integration

We have worked closely with CESD science domains, major DOE extreme-scale simulation groups, and DOE leadership-class computing facilities to identify key obstacles (such as integrating germane science analysis and visualization, data management, user support, data transportation, future data centers, etc.) to research productivity and scientific discovery in the integration of CESD's cyber-infrastructure environment. The overall integrated architecture will evolve proven existing CESD technologies and domain knowledge that are already in use by the broader community such as ESGF, ARM data services, and UV-CDAT [16] (see **Table 2 in Appendix 2**). The work will progress along two main directions: (1) extend and integrate the system to support all of

CESD's research activities and external collaborations, and (2) continue support data centers and other data intensive facilities to improve their data collection, processing, archival and distribution capabilities and facilitate further data integration (see **Figure 2**).

We envision that for each BER CESD scientific domain included in the BER VL, both data and metadata will be archived and accessed from distributed data centers. Each of these participating data centers will be part of one or more virtual scientific focus groups, thus allowing for sharing of data and metadata services with other data centers in the same scientific domain. A software stack will co-evolve to share data, metadata, data quality information, ontologies, visualization and analysis services between the data centers. In order to support a powerful and flexible access model, each service hosted on a data center will be exposed though a simple and well-documented service API (layered with security when appropriate) so that clients of different kinds can easily execute invocations and possibly chain requests in complex scientific workflows.

When needed, translation tools and middleware will be developed to interface between the data center services and the data and metadata collections at existing science data centers (see **Figure 2**). The design will also work with selected clients so that their core functionality can be easily invoked from a web environment or via command-line instructions with the appropriate arguments. This design will allow a close interaction between the web-hosted services and the tools installed on the user's desktop, which will access a vast amount of data and metadata at distributed locations as if it were local.

The architecture shown in **Figure 2** for the BER VL promotes the convergence of high-level service APIs towards discipline-neutral standards. For example, the same client can be used to search and download data from "Global Climate Modeling" or "Subsurface Biogeo-Chemistry" data centers. Whenever possible, the system will use or extend existing standards developed by the community, such as the OpenSearch specification for metadata querying and the Web Processing Service API for remote job execution. The general goal is to promote reuse of modular software components (described in **Table 2 in Appendix 2**) on the server- and client-side across multiple fields, while retaining attention to each BER CESD science domain's specific needs and requirements.

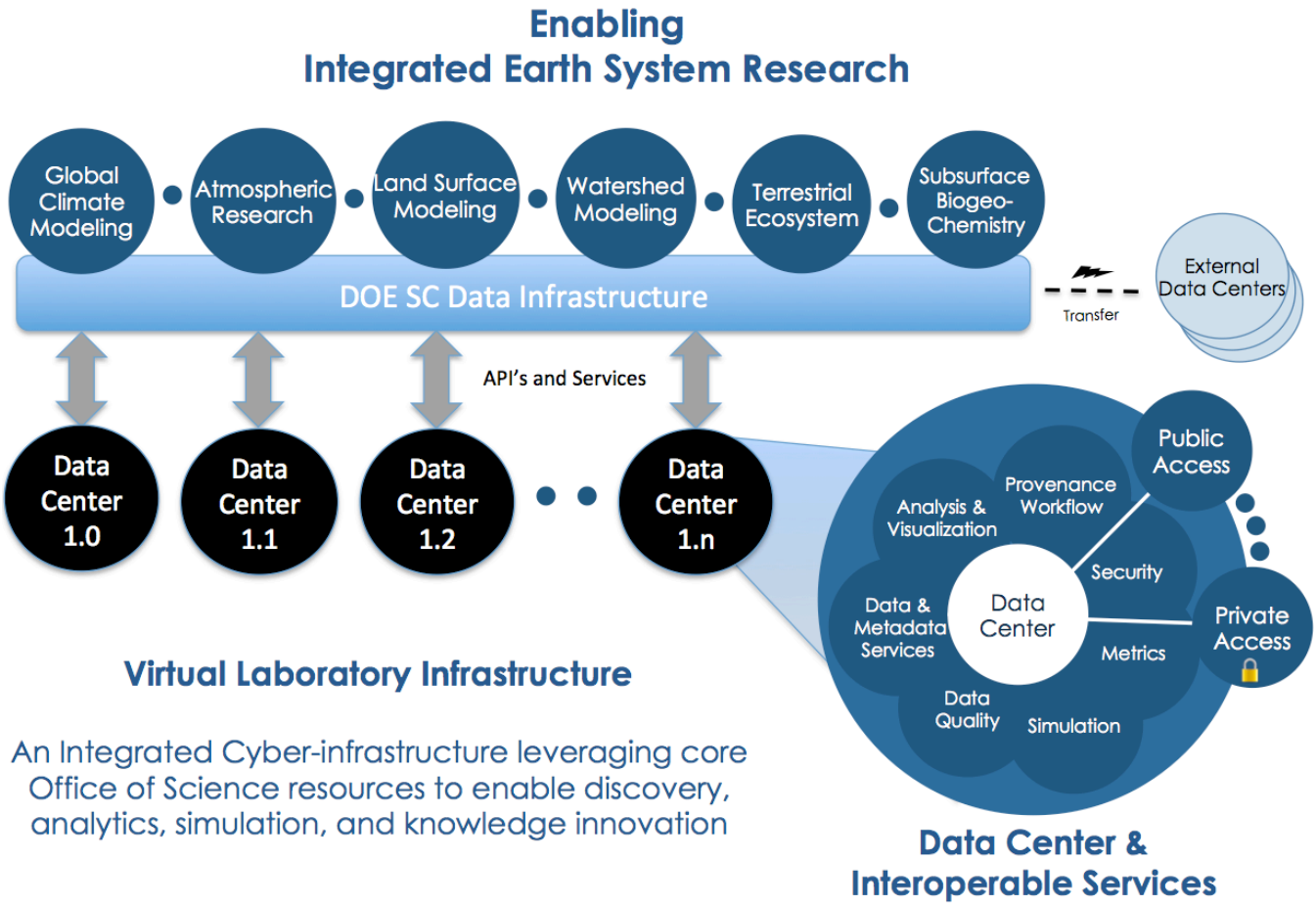

Figure 2. *The diagram depicts a scalable and flexible approach for managing CESD's data archives and services across ASCR and BER data centers and smaller facilities. Disseminating data, software, and computer services to the greater community, the integrated cyber-infrastructure, represents an environment of interoperable services designed to integrate diverse data holdings and process extreme scale data, heavily leveraging and advancing Office of Science resources.*

As an example strategy for ARM and CDIAC, participation in a federated data ecosystem such as ESGF is presented in **Figure 3**. The core concept here is to effectively use various interoperable standards, services, and tools to share data and additional information for innovative knowledge discovery.

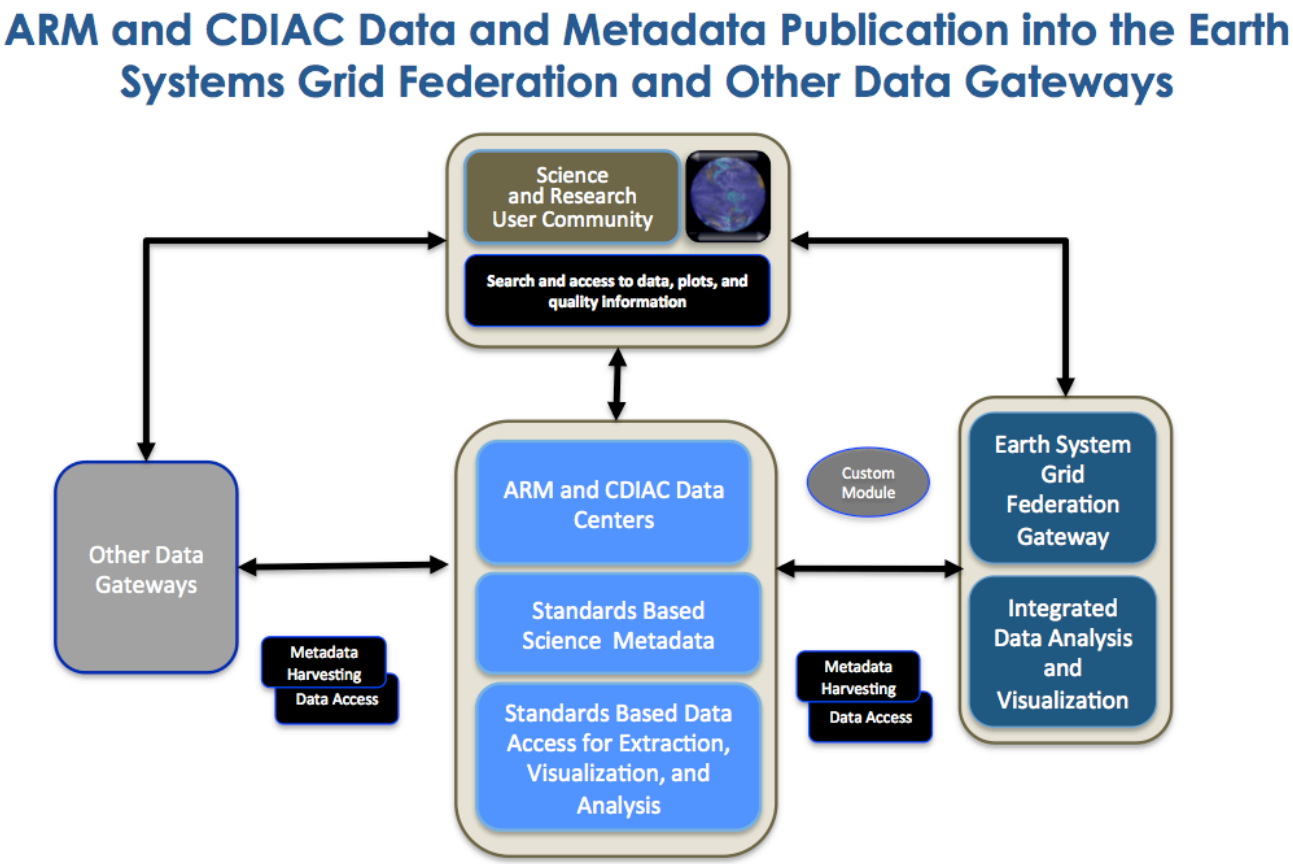

Figure 3. *The diagram depicts standards-based data sharing by BER CESD data centers (ARM and CDIAC) to community gateways and portals such as ESGF, US/EU Portal, UV-CDAT, and GEO.*

5.1 Data and metadata collection capabilities

DOE data centers have unique expertise in managing their data, and these capabilities will be preserved in the new architecture and possibly used by other upcoming data intensive projects. The BER VL proposed architecture will have an exemplary system for sharing these data and metadata records based on various community-developed standards such as ISO 19115, FGD, OAI-PMH, THREDDS and OGC. The architecture will also reuse some of the metadata creation tools such as OME, which is currently used by many DOE projects. Using OME will not only allow the users to register their data sets, but also to use consistent keywords using standards such as the Climate and Forecast- and Global Change Master Directory-controlled vocabularies.

The BER VL architecture will also have a common resource registry, which will allow the projects to register their resources such as tools, web services, and domain expertise using common protocols and standards.

5.2 Data quality

The importance of implementing a standards approach to understanding, communicating, documenting, and improving the quality of scientific data is paramount. Doing this job well is a fundamental step toward enabling an architecture for integrated Earth system science. Therefore, a key criterion in determining the utility of scientific data is quality. Data products used to advance the understanding of

complex interrelated research in modeling and fundamental physical processes must include a representation of quality in the data model and communicate that quality to the end user. Understanding data quality is a hierarchical process beginning with documenting systematic and random errors and the measurement environment. When working with scientific data across domains, this understanding is ever more important because users have to interpret and synthesize information for their analysis. Data quality services will need to be designed to generate a standard view of data product quality across BER data centers. The data quality service also needs to incorporate a feedback loop from the user back to the BER CESD science domains to document and resolve data quality deficiencies.

The formal terminology relating to measurements and uncertainties is set out in the International Vocabulary of Metrology (VIM) guidelines [17] and the "Guide to the expression of uncertainty in measurement" [18]. The proposed architecture will allow data centers to share the data quality information in a standards-based, machine-readable format, based on a common standard such as the ISO data quality extension [19].

5.3 Uncertainty quantification

Researchers test hypotheses that are framed and bounded by uncertainty. To gain or improve knowledge in any area of research, the data product uncertainty must be documented and the associated measurement errors should to be less than the anticipated uncertainties of the physical processes being analyzed. Problem solving in an integrated Earth system environment will require the generation of synthesis products, which involve the definition of relationships across physical measurements. Synthesis data products will have analytical and statistical methods applied across these relationships creating a calculus chain of uncertainties. The propagation of uncertainties has to be quantified and communicated to the scientific data user. There are standards-based approaches in place or under development that should be adopted to facilitate the communication of data product uncertainty of synthesis data products.

In addition, uncertainty quantification and data assimilation techniques are required to analytically model the subsystems of the end-to-end integrated data ecosystem (**Figure 1**) process and workflow and for predictive data infrastructure (**Figure 2**) behavior and corrections.

5.4 Ancillary information

This data roadmap does not focus solely on model output, primary observational measurements, and derived products. Ancillary data and information are essential to understanding and characterizing the BER CESD data collection and permitting this collection to be used to the fullest. Ancillary data and information, which help bound data types, prevent misuse or misapplication of data, inform models, permit synthesis studies, and advance science, include site characterizations, measurement heights and depths, sampling times, model parameterizations, carbon stock estimates, calibration standards, station histories and land-use histories, statistical summaries, web camera imagery, and so on. Ancillary data and information add richness and completeness to CESD's primary data holdings and expand the potential application of data generated by data centers.

5.5 Data preparation for archival

As synthesis data products are developed through analysis and research activities, the associated data product groups will be organized, registered, and archived for sharing within the CESD virtual collaboration environment. Requirements for standard file naming conventions, versioning, and provenance will be defined and implemented [20], as will rules to verify data integrity and consolidate

files into daily records for storage. An extensible database structure will be implemented and governed by a complete data model for the records and registration. This structure will be developed to facilitate exploration and relational associations across product holdings.

5.6 Data discovery and access

Collaboration and sharing policy: The open sharing of data among researchers, the broader scientific community, and the public is critical to meeting the scientific goals and objectives of BER CESD and critical to advancing the mission of DOE, the Office of Science (SC), and other BER directorates where the strategic intent is to deliver improved scientific data and models.

BER CESD data are freely available and are furnished by individual scientists who encourage their use or by projects committed to making data available to wider scientific audiences. This white paper recognizes and catalogs different levels of data maturity over time ranging from preliminary, proprietary, and project-exclusive data to public domain, open-access and secure data.

Users of BER CESD data are expected to inform data providers, whether the data originate from individual scientists or from central repositories, of how they intend to use the data and of any publication plans. These notification requirements are important to help assure users are downloading the latest revision of the data and to prevent potential misuse or misinterpretation of the data.

Users are expected to acknowledge the original data source as a citation in publications. Recommended data citations will be furnished to users as digital object identifiers assigned to data products and data streams. To foster collaboration, data contributors will be notified when fellow scientists retrieve their data.

Authentication and security: Data authentication and security architecture is essential to support the use cases mentioned in section 2, where the focus is on leveraging interoperable capabilities to examine constituent data in a regimented but free-flowing way. This will require modifications and extensions to exiting Earth system software security architecture [21, 22] that fit into BER CESD's overarching data strategic roadmap. In addition, the need to create a web of secure user identity formats and mapping for various resources, from services to HPC resources, will also need to be supported.

The integrated Earth system research execution platform will launch, monitor, control and specify how data processes are run and distributed across the data infrastructure services. The platform will also need to connect users and services. The execution platform creates an entirely new vehicle for sharing code among users in a secure way. It must be agnostic to the particular endpoint mechanisms and thus provide a machine-level abstraction of process execution in much the same way that it provides a unified platform for data management. It also requires linking, management and provisioning of multiple identity formats and types for a user, ranging from the user identity on the systems to where these executable programs are run at the data centers. The security architecture must be designed to support such a use case.

A number of security models today use a combination of Web single sign-on protocol (OpenID) and Public Key Infrastructure (X.509 Certificates) to provide authenticated access to data and other resources. In addition, these security models use the notion of groups to provide controlled access to data sets that need to be restricted to a subset of users. In these cases, the access control

policies heavily rely on groups and user roles. For the integrated Earth system, groups and roles need to be created and established in controlled name-spaces, to ensure secure use of key services and facilities across a wide variety of the supported cross-domain collaborations listed in **Figure 2**. By leveraging newer and simpler delegation protocols such as OAuth [23], in conjunction with Grid Security Infrastructure [24], this process will enable a powerful delegation mechanism with flexible features and audit capabilities, while still providing a streamlined user experience for the BER CESD data infrastructure user community.

Data analytical and visualization capabilities and services: Advanced tools for analyzing and visualizing ultra-scale Earth system data are required to maintain rapid progress in scientific understanding and prediction of climate change and its impacts and to apply it to the decision-making process. Analysis the visualization tools should be sufficiently flexible and scalable to incorporate existing and future software components with minimal or no infrastructure modifications, a feature that will allow the user to exploit other software packages normally not used for Earth science. In addition, the system must be modular so that the underlying technology is sustainable and applicable to other scientific disciplines, such as computational chemistry, biology, and informatics.

In general, data analytical and visualization capabilities and services must incorporate the following minimal requirements: interactive and batch operations; workflow analysis and provenance management; parallel visualization and analysis tools (exploiting parallel I/O); local and remote visualization and data access; comparative visualization and statistical analyses; robust tools for regridding, reprojection, and aggregation; and support for unstructured grids and non-gridded observational data, including geospatial formats often used for observational data sets. Parallel computing, exploratory analysis, big data processing for analysis, interactive analysis and visualization, and web-informatics are other key features that are also required.

Data downloading and subsetting services and capabilities: CESD data centers currently use various data downloading and subsetting services such as FTP download, Globus, Web and OGC services, and OPeNDAP. In addition, some have customized data subsetting and extraction capabilities as part of the data delivery options. The new architecture will identify and support sets of data downloading protocols, which could be used for effectively sharing the data.

User interface, portal (gateway), and APIs: DOE data centers currently use data portals customized to effectively serve their user community. The proposed architecture recognizes the need to preserve such customized portals, but it will also build common services and a higher-level data discovery portal and web services to discover and access data that are managed in distributed systems [25]. The web-based search system will allow a user to discover data using such search options as full-text search, fielded search, geospatial search, plot browsing, temporal search, and a hierarchical attribute browse search. The search results will contain the matched metadata records and an option to do faceted search refinements based on the logical grouping of various themes and keywords.

The proposed search tool will also have semantic-based data discovery, which allows the user to select related data sets. A higher level data search using semantics will resolve to finer variable-level search results, for example, users searching for “soil moisture” will also find other relevant metadata records that contain “soil water potential,” “soil moisture, gravimetric,” or “soil moisture, volumetric” as keywords. Various representations of variables can also be mapped

using semantics. For example, various representations of soil water potential will be mapped as a single variable, and unit conversion will be carried out using semantic relationships.

The proposed approach recognizes the importance of a service oriented architecture, hence it will allow users to query the services catalog using rich web user interface and other applications to query using web services, rich site summary (RSS), portlets, and other search sharing services. The proposed search APIs will also allow visualization and analysis tools to query the metadata index using standards based web service calls.

This infrastructure will be based on a rigorous API definition and on the adoption of industry standards such as SSL, PKI, OpenID and SAML. It will therefore allow clients and servers written in different languages to communicate with each other, within the CESD data infrastructure environment where access control is decentralized and managed by separate institutions. For example, a Python script running on a "Global Climate Modeling" machine can securely request data sets from a Java THREDDS Data Server at "Subsurface Biogeo-Chemistry," or a client application based on the NetCDF Java library running on a scientist's laptop can access data from other data centers.

The architecture will explore the possibilities of reusing the BER CESD Clearinghouse system as a data discovery tool, the BER CESD clearinghouse currently allows users to search data from four different data centers (ESGF, ARM, CDIAC, and NGEE and this could be expanded to support other DOE projects.

Ontology: A wide variety of data formats and different scientific communities use metadata standards. Users attempting to answer broader scientific questions such as factors inducing climate change require interdisciplinary data [26]. Proper ontology architecture is critical for discovering the data they need, confirming the usability, and integrating them in their analysis. The proposed architecture will effectively use community-developed ontologies, such as Semantic Web for Earth and Environmental Terminology (SWEET), OBOE (extensible observation ontology), ARM, ESG and CF-controlled vocabularies, to annotate the keywords found in the metadata records. These ontologies include several thousand terms, spanning a broad extent of Earth system science and related concepts. Using these ontologies will:

- Enhance the metadata keywords.
- Help users find various representations of a single parameter by providing variable level mappings, for example, air temperature could be defined as air_temp, atm_temp, air_temperature, all these could be mapped to a CF variable air_temperature.
- Help users to convert a measurement to a specific unit of their preference, for example, temperature measurement from K or °F to °C.

5.7 Integrating the data services

Data integration and advanced metadata capabilities: One important strategy to date has been to tightly specify metadata standards, which has proven successful. However, to accommodate a broader variety of environmental and scientific data, we will require a more extensible metadata infrastructure, more powerful processes for handling diverse data formats, and scalability up to billions of objects. We must develop mechanisms to describe and organize a wide variety of data not fully supported in any one community. In climate research, flexible data-format support will include mechanisms that allow users to work with complex data sets, such as long-term

measurements of carbon dioxide, water vapor, and energy exchanges representing different ecosystems in CESD (e.g., Ameriflux, NGEE, ARM, CMIP, see **Table 1 in Appendix 2**) and high-resolution global models, regional models, and observational data sets. The commonly used metadata standards include CF, ISO 19115, FGDC, and EML.

Performance of model execution: A significant challenge facing the climate science community is the extremely large data sets that are produced today from coupled model simulations and the projected increase in data set sizes as higher resolution models (T341) are commonly used. Scientists will require scalable tools and technologies for analysis of multi-terabyte data sets. These tools must be easy to use while scaling to HPC environments in which parallel analysis will be required due to processing time requirements and per-node memory capacity. Services today are emerging that couple the capabilities of HPC environments with web-based service delivery mechanisms. These multi-tiered "applications" provide users with access to high performance parallel analysis routines hosted on HPC platforms using common technologies such as MPI and OpenMP coupled with a web-services framework using standard RESTful interfaces. We envision building upon and extending this approach by providing a rich ecosystem of analysis services across DOE SC compute and data resources and an equally rich ecosystem of user interfaces/views. Recognizing that many users will require access to these services within a terminal environment (command line), we envision making these same analysis services available in a rich scripting environment.

Analysis services when multiple data sets are not co-located: To enable truly integrative research, many analysis tasks will need to fuse data sets from multiple data centers. In some cases, this analysis can be accomplished by running the constituent parts of the analysis on computational infrastructure co-resident with the data. An example of this type of analysis would be calculating an average temperature over a given set of months at a specific grid cell across multiple climate model outputs hosted at different data centers. This task can easily be distributed across the computational infrastructure at the data centers and then aggregated back at the original computational resource. This example requires very little data movement, and integration of the resultant analysis can easily be accomplished on a typical workstation. In other cases, analysis tasks will require access to data across centers concurrently.

Advanced product services (i.e., exploratory, specialized, etc.): The product services provide users with custom visualization, subsetting, and basic analysis and exploratory capabilities applied to the underlying data collection via a browser-based interface or via command-line. These services are essential to serving a diverse user community. It is essential to keep in mind scientists not accustomed to working with complex output domains and who need to quickly discern which data are suitable for their needs. In the BER VL, non-HPC experts will have ease of access to pre-built analysis products based on both smaller scale as well as extreme scale datasets.

A dense observation such as on-the-fly objective analysis (interpolations) of observations superimposed on compatible grids is another example of an advanced product needed for the integration of services.

Advanced networks as easy-to-use community resources: A primary goal of the BER VL will be to provide BER CESD scientists with tools to manage and analyze extreme-scale Earth system data using the first-ever ESnet 100-Gbps backbone [27] and large- and mid-range LCF and NERSC computing resources. Scientists need advanced tools to apply these resources effectively so they can manage and manipulate data from trivial to extreme scales. DOE, and indeed the world

community, is making significant investments in hardware, network, and software services and resources; however, the researchers and non-researchers do not yet know how best to work with them. A goal should be to ensure that all services and resources are easy-to-use by all and have needed documentation accessible for users.

5.8 Provenance and workflow

A workflow and provenance environment is needed to easily reproduce analyses and other products for anyone requesting the results of work entered into the BER data infrastructure. Support for enabling integrated Earth system research will therefore include the delivery of ultra-large data and diagnostic products to collaborating scientists (and quite possibly others not familiar with the field). In addition to making document workflows and provenance more transparent, common workflows can also be used to optimize access to the data and the analyses involved, reducing the data transmission load on the infrastructure. The broad community of researchers will be able to efficiently access the most popular data products and trace how they were produced in a repeatable fashion.

Provenance capture throughout the infrastructure containing model runs, diagnostics, production cycles, and rich research will enable unprecedented levels of reproducibility. For scientific reproducibility and collaboration, remote services, including analysis and visualization, will be connected through a provenance API to automatically capture meaningful history and workflow. Two examples include sharing this captured information with others or using it to reproduce a set of events at a later date.

The data centers are standalone instances representing various research areas shown in **Figure 2** as well as being part of a larger ecosystem that allows users to run models and collect workflow and provenance independently for sharing scientific results and enhancing reproducibility.

5.9 Compute and data services

The integrated cyber-infrastructure (**Figure 2**) will build upon core compute and data services supported by the DOE SC. These facilities include large-scale compute and data storage resources available at NERSC and the LCFs at Argonne and Oak Ridge, advanced networking infrastructure from ESnet, CESD data lifecycle management capabilities at ARM, CDIAC, NGEE-Arctic and CMIP, and emerging data services in the SC.

To support integration and intercomparison of the increasing data volumes generated across CESD, from observation networks, to modeling activities, large-scale data storage infrastructure will play a significant role in the integrated cyber-infrastructure. Today, high-resolution climate modeling activities are generating tens to hundreds of terabytes of data for a single study and are projected to generate tens to hundreds of petabytes of data over the next decade. Observational capabilities such as ARM's scanning cloud radar are capable of generating multi-terabyte data sets today, and over the life of these systems will generate many petabytes. Simply storing these data sets will be a challenge while providing high-performance compute infrastructure—allowing efficient retrieval and analysis is an even greater one. Meeting the growing demand for storage and compute infrastructure will necessitate a highly leveraged approach, one which makes judicious use of exiting BER data resources more closely coupled with compute and data infrastructure provided at the ASCR facilities.

Leveraging the computational and data fabric of DOE SC will rely upon the advanced networking capabilities of ESnet. For large-scale data replication tasks (large flows) the use of OSCARS provisioning would enable deterministic bandwidth across ESnet between the cyber-infrastructure allowing high-performance data motion while minimizing the impact to end users (smaller flows). This

will require data replication and transfer software stacks to take advantage of these advanced networking capabilities, insulating the user community from the nuances of using these services.

As DOE SC continues to expand the set of available compute and data services, we will incorporate these capabilities into the BER VL and build upon them to deliver end solutions to BER CESD. We envision that DOE SC will increasingly move towards providing robust Infrastructure-as-a-Service (IaaS), our software development and integration will therefore target these technologies, allowing broad adoption of our software. Similarly, should DOE SC begin to deploy Software-as-a-Service (SaaS) offerings such as Map-Reduce or NoSQL storage as a service, we will rapidly incorporate these technologies within our software stack.

5.10 Dashboard and system monitoring

User and data product usage metrics and reporting: Data product generation of the SC national user facilities and associated research components have processes and rules for registering scientific users and recording their requests. These processes are designed to create a cooperative and consistent view of user identity, information, research purpose, data product requests, and demographics. Metrics for the national scientific facilities of the SC track and report quarterly "unique scientific users" by number and affiliation and the data product usage patterns are available for analysis. The virtual collaboration environment for BER data centers will be designed to incorporate SC user and data product usage rules and processes, and, provide user and data product usage statistics feedback to data centers served through the virtual environment.

Data access and pattern tracking: As users interact with the BER VL, a record of queries will be inventoried for analysis in conjunction with the products delivered to the user. The associated number of files and bytes of all transactions will be joined through a data product delivery service. This information will provide valuable relationships and research insights on the products that are being requested for scientific research. Also, this information can be used to facilitate product relevance across disciplines and provide an index back to users if prior product states change.

Network monitoring (tracking network speeds and usability): The DOE SC has built a national-scale, 100 Gb/s networking facility (ESnet) dedicated to improving the scientific productivity of researchers working in areas of national strategic importance such as biology, climate change, and energy. As DOE scientists and collaborators worldwide have long required high-bandwidth direct connections among major DOE sites and with the international community, the emergence of distributed large-scale science demands a network fast enough to meet the needs of the new, highly distributed model. With this, comes the need to monitor the network speeds and its usability. With the help of ESnet and other network resources and expertise, the movement of data will be tracked for performance value and for greater acceleration of scientific knowledge. Achieving this capability on production systems will help to prepare the BER CESD infrastructure for the demands of future large-scale data activities such as CMIP6, ARM, etc. and set the stage for continued scientific productivity in other critically important areas of BER CESD Earth system science.

6 Help and Support

To ensure effective implementation and control of training activities and user support for the BER VL, a community help desk must be established to assist users with information and resources and to help

troubleshoot problems with the BER VL integrated data ecosystem, data, analysis, visualization and other products. This help desk service is provided in the form of software, website, and e-mail. Traffic on the help desk will be monitored to inform whether or not adequate support is meeting usage volumes for the BER VL integrated ecosystem. The development team members, along with scientists and general users of the system, will be essential to keeping this operation going. That is, from experience with other existing projects, we know that help desk traffic is considerable, requiring user responses back and forth with help desk providers and follow-on questions by other users.

Scientists within the BER VL will be tasked with addressing data and science questions, while technical staff at data centers will be charged with addressing system questions. These personnel can register for specific questions and spend a portion of their day scanning the list of new questions, divvying up the workload, and responding to users. Questions that are resolved should be placed on a frequently asked question (FAQ) list to circumvent answering repeat questions. This work includes the training necessary for new projects, data providers and the diverse user community. It also includes: software-use documentation and tutorials; support mailing lists; and websites and wiki sites.

7 Operational and Modernization

The BER VL integrated data infrastructure will never be a static system. As the platforms on which it operates—server hardware, networks, operating systems, and browsers—evolve, the data centers and interoperable services must be adapted. The infrastructure will also be a collaborative development with components from several quasi-independent projects shown in **Table 2 in Appendix 2**. As on component advances or modernizes, adaptations in others are inevitable, despite best attempts to isolate functionality through interface definitions. Therefore, additional operational and maintenance support will need to be continued for the DOE BER CESD data infrastructure. This work is to include: data center hardware (e.g., deep storage, computer clusters, etc.) and network maintenance and refresh; data management and transport; analysis and visualization updates; user interface development and maintenance; security; publishing; product services; and overall user support.

8 Summary

A longstanding challenge for scientists in the BER CESD Earth science community has been the difficulty of incorporating, inspecting, and analyzing data with newly developed technologies, diagnostics, and visualization techniques in an efficient and flexible way. To improve research ability and productivity, an integrated data infrastructure must be in place to help make vital and quick strategic decisions reflecting the future of Earth's climate and energy. To help meet this challenges, the authors of this white paper are proposing the establishment of a BER Virtual Laboratory (**Figure 2**) to help integrate disparate community software tools for the discovery, examination, and intercomparison of coupled multi-model and observational climate data sets. The BER Virtual Laboratory will bring together top climate institutions, computational organizations, and other science communities to provide proven data management, analysis, visualization, diagnostics, network, and hardware capabilities to BER CESD scientists. This effort is focused on developing several powerful and insightful applications for knowledge discovery of observed and simulation climate data and orchestrate these activities into several national and international organizations dedicated to improving the climate science data infrastructure.

Despite the many attractive features of the BER VL's proposed data management and analysis infrastructure, several challenges and data science research issues remain. Enabling meaningful and

credible integration of data types across varying spatial and temporal domains within the participating science projects and data centers will be a challenge. Integrating data from multiple sources for users in a transparent fashion, both within the BER CESD data infrastructure and outside, through the proposed higher-level data portal while maintaining proper source attribution and provenance will need to be addressed.

9 References

[1] Janet Braam, Judith A. Curry, et al., BER Virtual Laboratory: Innovative Framework for Biological and Environmental Grand Challenges. Office of Biological and Environmental Research, Office of Science, Department of Energy, http://genomicscience.energy.gov/program/beracvirtuallab.shtml.

[2] Overpeck, J.T., G. A. Meehl, S. Bony, and D. R. Easterling, 2011: Climate Data Challenges in the 21st Century. Science, vol. 331, no. 6018, pp. 700-702, dx.doi.org/10.1126/science.1197869.

[3] The Earth System Grid Federation home page. http://esgf.org/.

[4] Carbon Dioxide Information Analysis Center home page. http://cdiac.esd.ornl.gov/.

[5] Luca Cinquini, Daniel Crichton, Chris Mattmann, Gavin M. Bell, Bob Drach, Dean Williams, John Harney, Galen Shipman, Feiyi Wang, Philip Kershaw, Stephen Pascoe, Rachana Ananthakrishnan, Neill Miller, Estanislao Gonzalez, Sebastian Denvil, Mark Morgan, Sandro Fiore, Zed Pobre, Roland Schweitzer, "The Earth System Grid Federation: An Open Infrastructure for Access to Distributed Geospatial Data", IEEE Future Generation Computer Systems, http://dx.doi.org/10.1016/j.future.2013.07.002, 17 September. 2013.

[6] ARM Climate Research Facility home page. http://www.arm.gov/.

[7] Interactive Web-based tool for viewing Atmospheric Radiation Measurement (ARM) data website [Accessed April 2014]; Available from: https://ams.confex.com/ams/annual2003/techprogram/paper_55288.htm.

[8] DOE ASCAC Data Subcommittee Report, "Synergistic Challenges in Data-Intensive Science and Exascale Computing," technical Report, U.S. Department of Energy Office of Science, March 2013.

[9] National Science Foundation Advisory Committee Task Force on Data and visualization: https://www.nsf.gov/cise/aci/taskforces/TaskForceReport_Data.pdf .

[10] Biological and Environmental Research Reports and Workshops Series. [Accessed April 2014]; Available from: http://science.energy.gov/ber/news-and-resources/.

[11] Basic Energy Sciences Report Series [Accessed April 2014]; Available from: http://science.energy.gov/bes/news-and-resources/reports/.

[12] Earth System Grid Federation and Ultra Visualization Climate Data Analysis Tools Face-to- Face Meeting, http://aims-group.github.io/pdf/ESGF_UV- CDAT_Meeting_Report_March2014.pdf.

[13] Biological and Environmental Research Reports and Workshops Series. [Accessed April 2014]; Available from: http://science.energy.gov/ber/news-and-resources/.

[14] Obama Administration Announces "Big Data" Initiative: Unveils $200M in New Research Investments, http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf, 2012.

[15] Mount, R. and Skinner, D. (eds.). Scientific Collaborations for Extreme-Scale Science. Office of Advanced Scientific Computing Research, Office of Science, Department of Energy, http://bit.ly/JcHxIA 2011.

[16] Dean N. Williams, Timo Bremer, Charles Doutriaux, John Patchett, Sean Williams, Galen Shipman, Ross Miller, David R. Pugmire, Brian Smith, Chad Steed, E. Wes Bethel, Hank Childs, Harinarayan Krishnan, Prabhat, Michael Wehner, Claudio T. Silva, Emanuele Santos, David Koop, Tommy Ellqvist, Jorge Poco, Berk Geveci, Aashish Chaudhary, Andy Bauer, Alexander Pletzer, David Kindig, Gerald L. Potter, Thomas P. Maxwell, "Ultrascale Visualization of Climate Data", IEEE Computer Magazine, September 2013, vol. 46 no 9, pp. 68-76.

[17] Joint Committee for Guides in Metrology (VIM) home page. http://www.iso.org/sites/JCGM/VIM-introduction.htm

[18] Joint Committee for Guides in Metrology (JCGM) Working Group (GUM) home page. http://www.iso.org/sites/JCGM/GUM-introduction.htm

[19] NetCDF-U the NetCDF Extension for Uncertainty, Uncert|Web, UncertML, ISO 19113 Quality Principles for Geographic Information, ISO 19139 Metadata and XML Schema Implementation for Geographic Information, ISO 19157 Data Quality for Geographic Information, ISO 19115 Metadata for Geographic Information.

[20] The Federal Geographic Data Committee Geospatial Metadata Standards home page. http://www.fgdc.gov/metadata/geospatial-metadata-standards.

[21] Siebenlist, F., Ananthakrishnan, R., Foster, I., Miller, N., Bernholdt, D., Cinquini, L., Middleton, D.E. and Williams, D.N., The Earth System Grid Authentication Infrastructure: Integrating Local Authentication, OpenID and PKI. TeraGrid '09, Arlington, VA, 2009.

[22] Foster, I., Kesselman, C., Tsudik, G. and Tuecke, S., A Security Architecture for Computational Grids. 5th ACM Conference on Computer and Communications Security, 1998, 83-91.

[23] Ronan-Alexandre Cherrueau, Rémi Douence, Jean-Claude Royer, Mario Südholt, Anderson Santana de Oliveira, Yves Roudier, Matteo Dell'Amico. Reference Monitors for Security and Interoperabiltiy in OAuth 2.0. Conference: Data Privacy Management and Autonomous Spontaneous Security – 8th International Workshop, Egham, UK, 2013, pp. 235-249, dx.doi.org/10.1007/978-3-642-54568-9_15.

[24] Peter Tröger, Andre Merzky, Towards Standardized Job Submission and Control in Infrastruture Clouds. Journal of Grid Computing, March 2014, vol. 12, issue 1, pp 111-125.

[25] Altinay, C., Binsteiner, M., Gross, L. and Weatherley, D.K., High-Performance Scientific Computing for the Masses: Developing Secure Grid Portals for Scientific Workflows. e-Science (e-Science), 2010 IEEE Sixth International Conference on, 2010, 254-260.

[26] Pouchard, LineC., MarciaL. Branstetter, RobertB. Cook, Ranjeet Devarakonda, Jim Green, Giri Palanisamy, Paul Alexander, and NatalyaF. Noy. "A Linked Science Investigation: Enhancing Climate Change Data Discovery with Semantic Technologies." Earth Science Informatics 6, no. 3 (September 1, 2013): 175–85. http://dx.doi.org/10.1007/s12145-013-0118-2.

[27] Eli Dart, Brian Tierney, Editors, "Biological and Environmental Research Network Requirements Workshop, November 2012 - Final Report", November 29, 2012, LBNL LBNL-6395E http://www.es.net/assets/pubs_presos/BER-Net-Req-Review-2012-Final-Report.pdf.

Appendix 1: Glossary of Terms

Acronym	Meaning and Website
ACME	Accelerated Climate Modeling for Energy: DOE's effort to build an Earth system modeling capability tailored to meet the climate change research strategic objectives
API	Application Programming Interface (http://en.wikipedia.org/wiki/Application_programming_interface)
AR5	Fifth IPCC Assessment Report, published in 2013 (http://www.ipcc.ch/report/ar5/#.UwVGOCTm6Gg)
ARM	Atmospheric Radiation Measurement is a U.S. Department of Energy scientific user facility, providing data from strategically located in situ and remote sensing observatories around the world (http://www.arm.gov/)
BER	Office of Biological and Environmental Research under the DOE Office of Science (http://science.energy.gov/ber/)
CDIAC	Carbon Dioxide Information Analysis Center (http://cdiac.esd.ornl.gov/)
CESD	Climate and Environmental Sciences Division under DOE's Office of Biological and Environmental Research (http://science.energy.gov/ber/research/cesd/)
CF	Climate and Forecast metadata convention, for processing and sharing NetCDF data files (http://cf-pcmdi.llnl.gov/)
Client-Server	Relationship between two computer programs, where the client program makes a service request, which the server program fulfills (http://en.wikipedia.org/wiki/Client-server)
CMIP5	Coupled Model Intercomparison Project 5, sponsored by WCRP/WGCM, and related multi-model database planned for the IPCC AR5 (http://cmip-pcmdi.llnl.gov)
CMIP6	Coupled Model Intercomparison Project 6, sponsored by WCRP/WGCM, and related multi-model database planned for the IPCC AR6 (http://www.wcrp-climate.org/index.php/wgcm-cmip/wgcm-cmip6)
Data Center	A facility used to house computer systems and associated components, such as HPCs, clusters, storage systems, communications, etc. (http://en.wikipedia.org/wiki/Data_center)
DOE	Department of Energy, the U.S. government entity chiefly responsible for implementing energy policy (http://www.doe.gov/)
EDEN	Exploratory Data analysis Environment is a visual analytics tool for exploring multivariate data sets. (http://cda.ornl.gov/projects/eden/)
ESGF	Earth System Grid Federation, led by LLNL, a worldwide federation of climate and computer scientists deploying a distributed multi-petabyte archive for climate science (http://esgf.org)
Esnet	DOE Energy Science Network (https://www.es.net/)
FACE	Free-Air Carbon dioxide Enrichment
FAQs	Frequently Asked Questions

Gbps	Gigabit per second, 10^9 bits of information (http://en.wikipedia.org/wiki/Data_rate_units)
GCMD	Global Change Master Directory, a directory of descriptions of data sets of relevance to global change research (http://gcmd.nasa.gov/)
GEO	Group on Earth Observations (http://www.Earthobservation.org/index.shtml)
Globus	Globus is an open-source software toolkit used for building grids (https://www.globus.org/)
GridFTP	Is an extension of the standard File Transfer Protocol for high-speed, reliable, and secure data transfer (http://toolkit.globus.org/toolkit/docs/latest-stable/gridftp/)
HIPPO	HIAPER Pole-to-Pole Observations is a Java based open source Web content management system platform (http://hippo.ornl.gov/)
HPC	High-Performance Computing (http://en.wikipedia.org/wiki/Supercomputer)
IaaS	Infrastructure as a Service, a provision model in which an organization outsources the equipment used to support operations, including storage, hardware, severs and networking components (http://www.gartner.com/it-glossary/infrastructure-as-a-service-iaas)
ILAMB	The International Land Model Benchmarking project (http://www.ilamb.org/about/contacts.html)
IPCC	Intergovernmental Panel on Climate Change, a scientific body of the United Nations, periodically issues assessment reports on climate change (http://www.ipcc.ch/)
ISO	International Organization for Standardization, an international standard-setting body composed of representatives from various national standards organizations (http://www.iso.org/iso/home.html)
LCF	Leadership Computing Facility, DOE has two LCFs, one at Oak Ridge (OLCF – https://www.olcf.ornl.gov/) and other at Argonne (ALCF – https://www.alcf.anl.gov/)
LLNL	Lawrence Livermore National Laboratory, sponsored by the DOE (https://www.llnl.gov/)
Metadata	Data properties, such as their origins, spatio-temporal extent, and format (http://en.wikipedia.org/wiki/Metadata)
MIPs	Model Intercomparison Projects. There are over 70 MIPs worldwide. An example of a MIP is obs4MIPS, an activity to make observational products more accessible for climate model intercomparisons (i.e., CMIP) (http://obs4mips.llnl.gov:8080/wiki/)
NetCDF	A machine-independent, self-describing, binary data format (http://www.unidata.ucar.edu/software/netcdf/)
NERSC	National Energy Research Scientific Computing Center (https://www.nersc.gov/)
NGEE	Next-Generation Ecosystem Experiments (http://ngee-arctic.ornl.gov/)
OBOE	Extensible Observation Ontology, an otology for ecological observational data (https://marinemetadata.org/references/oboeontology)
OGC	Open Geospatial Consortium is an international voluntary consensus standard □rganization (http://www.opengeospatial.org/)

OMG	Online Metadata Editor, a tool to help document science data (http://mercury-ops2.ornl.gov/OME/)
OPeNDAP	Open-source Project for Network Data Access Protocol is a data transport architecture and protocol widely used by Earth scientists (http://www.opendap.org/)
OpenID	Allows users to use an existing account to sign in to multiple websites, without needing to create new passwords (http://openid.net/)
ORNL	Oak Ridge National Laboratory, sponsored by the DOE (http://www.ornl.gov/)
PB	Petabyte, 10^{15} bytes of information (http://en.wikipedia.org/wiki/Petabyte)
PCMDI	Program for Climate Model Diagnosis and Intercomparison, located at LLNL (http://www-pcmdi.llnl.gov/)
PKI	Public Key Infrastructure is a set of hardware, software, people, policies, and procedures needed to create, manage, distribute, use, store, and revoke digital certificates (http://en.wikipedia.org/wiki/Public-key_infrastructure)
REST	Representational State Transfer, a software architectural style consisting of a coordinated set of architectural constraints applied to components, connectors and data elements, within a distributed system (http://en.wikipedia.org/wiki/Representational_state_transfer)
SaaS	Software as a Service a software delivery method that provides access to software and its functions remotely as a Web-based service (http://www.gartner.com/it-glossary/infrastructure-as-a-service-iaas)
SC	Office of Science, a Program Office within the Department of Energy (http://science.energy.gov/)
SSL	Secure Sockets Layer (http://en.wikipedia.org/wiki/Secure_Sockets_Layer)
THREDDS	Thematic Real-time Environmental Distributed Data Services (https://www.unidata.ucar.edu/software/thredds/current/tds/)
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 (http://uv-cdat.org)
VIM	International Vocabulary of Metrology (http://www.bipm.org/en/publications/guides/vim.html)
VL	Virtual Laboratory is an interactive environment for creating and conducting simulated experiments. It consists of domain-dependent simulation programs, experimental units called objects that encompass data files, tools, services, resources that operate on these objects (http://edutechwiki.unige.ch/en/Virtual_laboratory)
WCRP	World Climate Research Programme, which aims to facilitate analysis and prediction of Earth system variability and change for use in an increasing range of practical applications of direct relevance, benefit, and value to society (http://www.wcrp-climate.org/)
WGCM	Working Group on Coupled Modeling (http://www.wcrp-climate.org/wgcm/)
Web portal	A point of access to information on the World Wide Web (http://en.wikipedia.org/wiki/Web_portal)

Appendix 2: Tables of data driven CESD projects, tools, and services

Table 1. Inventory of existing CESD projects that generate and rely on data.

Data generating projects and holdings	Point of Contact	Description
ACME	Dave Bader (proxy Dean N. Williams	Gridded and station data equivalents. High and lower temporal resolution
AmeriFlux	Margaret Torn/Tom Boden	AmeriFlux is a network of ~150 past and present terrestrial sites in Central, North, and South America making continuous measurements of carbon dioxide, radiation, and water vapor fluxes along with continuous meteorological measurements. Station data with various data versions and product levels. Station data at multiple heights and depths. Measurement frequencies vary from continuous eddy-covariance measurements to infrequent biological samplings.
ARM	Giri Palanisamy/Jimmy Voyles	The ARM Data Archive collects and distributes over 4000 ARM observational and PI data products. The ARM data management includes data collection and preparation, data quality, data and metadata dissemination services, metrics collection and reporting, data processing as-a-service. In addition, ARM has a well defined metadata architecture which is used in data discovery and access.
CAPT	Shaocheng Xie	Gridded forcing data are high-resolution analyses. Outputs from models are either limited spatial domain regions or station data-like output.
CASCADE	?	High resolution model runs, data are gridded
CDIAC	Tom Boden	CDIAC's data collection covers numerous disciplines (chemical oceanography, meteorology, climatology, atmospheric chemistry), time periods, and geographic representations (global to microenvironments). The collection includes original data and derived, value-added data products.
CMIP (PCMDI)	Karl Taylor (proxy Dean N. Williams)	Gridded GCM data. Many temporal resolutions.
COSIM	?	Gridded regular and irregular grids, transects, in-situ data and high temporal frequency
EMSL	?	Cell isolation and system analysis, deposition and micro-fabrication, mass spectrometry, microscopy and molecular science computing
FACE	Tom Boden/Rich Norby	Station data at multiple heights and depths. Measurement frequencies vary from continuous measurements to infrequent biological samplings.
ILAMB	Forrest Hoffman	Gridded data as well as station data, high temporal frequency.
NGEE Arctic Data Archive	Tom Boden/Giri Palanisamy/Stan Wullschleger	Measurements are made and samples are collected at multiple heights and depths. Measurement frequencies vary from continuous measurements to infrequent biological samplings. Station data and merged products.
Obs4MIPs and 70 other model intercomparison projects (MIPs)	Dean N. Williams	Gridded data of comparable form to CMIP efforts.

RCM	?	High resolution in space and time gridded data, some station data equivalents are output for direct comparison to ARM data.
SPRUCE Data Archive	Les Hook/Misha Krassovski/Tom Boden/Paul Hanson	Measurements are made and samples are collected at multiple heights and depths. Measurement frequencies vary from continuous measurements to infrequent biological samplings. Station data and merged products.

Table 2. Inventory of existing CESD data tools and services.

Related Data Projects	Point of Contact	Description
Accelerated Climate Modeling for Energy (ACME) Test Bed and Workflow Framework	Dean N. Williams	The test bed environment will provide the group of collaborating DOE scientists with the data and computing infrastructure needed for rapid development and assessment of new scientific modules and provide a testing-to-production environment for simulation and evaluation (i.e., metrics, diagnosis, and intercomparison).
AmeriFlux Network Management Center/ AmerFlux Site and Data Exploration System	Margaret Torn/Tom Boden	The AmeriFlux Network Management Center coordinates measurement activities at core AmeriFlux sites , handles data submissions and processing, and creates value-added data products (http://ameriflx.lb.gov). Processed data may be queried , viewed,.and retrieved through a data interface.(http://ameriflux.ornl.gov).
ARM Data Services	Giri Palanisamy/Jimmy Voyles	ARM has many data related services including: Web based ARM data visualization tools. Machine Readable Data Quality web services, data change notification workflow, metadata and data web services etc.
Atmospheric Radiation Measurement Data Integration.(ADI) Service	Jimmy Voyles	ADI allows ARM users and infrastructure to help prepare data using common set of tools and services. It also helps PIs to generate value added data product generations.
ARM Radar Data Processing Cluster Ecosystem	Giri Palanisamy/Raymond McCord	ARM radar cluster allows ARM users to easily access and analyze large amounts of radar data using a variety of tools and services. The cluster has customized tools including IDL, MatLab, ARM netCDF viewer, SciPy, TITAN and ADI.
BER CESD Data Clearinghouse (Mercury)	Giri Palanisamy/Tom Boden	Clearinghouse harvests metadata and ancillary.data from desperate data centers and sources and builds a central index for next generation scientific discovery with advanced search capabilities. It uses various community metadata standards including FGDC, ISO, CF, EML etc. It also has a n ontology-based search for cross domain data discovery.
Climate and Forecast (CF) and the Climate Model Output Rewriter (CMOR)	Karl Taylor (proxy Dean N. Williams)	Metadata conventions that designed to promote the processing and sharing of files created with the netCDF API. The conventions define metadata that provide a definitive description of what the data in each variable represents, and the spatial and temporal properties of the data. This enables users of data from different sources to decide which quantities are comparable, and facilitates building applications with powerful extraction, regridding, and display capabilities. http://cf-convention.github.io/
Carbon Dioxide	Tom Boden	CDIAC serves as a central repository for a broad array of

Information Analysis Center (CDIAC) Data Servies		climate-related data generated worldwide (http://cdiac.ornl.gov). Numerous project-level databases reside at CDIAC and are available through user interfaces.(e.g., Oceans CO2.@ http://cdiac.ornl.gov/oceans/, SPRUCE @ http://mnspruce.ornl.gov/)
Earth System Grid Federation (ESGF)	Dean N. Williams	ESGF is a coordinated multi-agency, international collaboration of institutions that continually develop, deploy, and maintain software needed to facilitate and empower the study of climate change. Through ESGF, users access, analyze, and visualize data using a globally federated collection of networks, computers, and software. http://esgf.org
Exploratory Data Analysis Environment (EDEN)	Galen Shipman / Chad Steed	EDEN is a visual analytics tool for exploring multivariate data sets. EDEN helps you see the associations among variables for guided analysis. EDEN has been used by Climate Scientists at ORNL and NCAR for land model analytics. http://cda.ornl.gov/projects/eden/
International Climate Network Working Group (ICNWG)	Dean N. Williams	ICNWG is a collaboration of international network organizations in Australia (AARnet), Germany (DFN), the Netherlands (SURFnet), the UK (Janet), and the US (ESnet). It is helping to set up and optimize network infrastructure for multiple climate data sites located throughout DOE and around the world. Through ICNWG, climate and computational scientists manage and disseminate petabytes of modeling and observational data, which traverse more than 13,000 miles of networks, spanning two oceans, and three continents. http://icnwg.llnl.gov/
PI Data publication Tool (Online Metadata Editor)	Giri Palanisamy/Tom Boden	The workflow for principal investigators to submit ARM science research products, field campaign data, or DOE-supported research data to the ARM Data Archive using a standards-based online metadata editor. The customizable OME is currently supporting nearly a dozen data centers/projects including ARM, CDIAC, NGEE, OCEAN, US/EU etc.
Parallel Climate Analysis Tool (ParCAT)	Galen Shipman / Brian Smith	ParCAT (parallel climate analysis tool) is a focused, easy-to-use parallel analysis tool primarily for climate data analysis. ParCAT parallelizes many common tasks in climate data analysis -- model diagnostics and verification, ensemble run comparisons, and visualization.
Ultra-scale Visualization Climate Data Analysis Tools (UV-CDAT)	Dean N. Williams	UV-CDAT is an international multi-institution open-source software collaboration that brings together disparate software subsystems and packages to form an integrated environment for data analysis and model diagnosis. It provides workflow, provenance, and scalable visualization techniques to a large community of users. http://uv-cdat.org

Table 3. Large scale computing facilities and network providers.

Hardware and Network	Point of Contact	Description
OCLF & CADES	Galen Shipman	The Oak Ridge Leadership Computing Facility mission is accelerating scientific discovery and engineering progress by providing outstanding computing and data management resources to high-priority research and development projects. Key resources within the OLCF include Titan, the 27 Petaflop supercomputer, Rhea a mid-scale data analysis cluster, and large-scale storage infrastructure. The Compute and Data Environment for Science at Oak Ridge National Laboratory provides R&D programs with with a customizable compute and data environment for a variety of use cases including large-scale data archives, data capture services suitable for experiment/observation facilities, workflow infrastructure, interactive analytics, and the ability to integrate with other computational and data resources across DOE SC. Key resources include compute and storage systems suitable for data portals, workflow systems, and interactive analysis platforms.
ACLF	Douglas Waldron	The ALCF's mission is to accelerate major scientific discoveries and engineering breakthroughs for humanity by designing and providing world-leading computing facilities in partnership with the computational science community. Key resources include Mira, the 10 Petaflop supercomputer, Tukey, a mid-scale data analysis cluster, and large-scale storage infrastructure.
NERSC	David Skinner	The mission of the National Energy Research Scientific Computing Center (NERSC) is to accelerate scientific discovery at the DOE Office of Science through high performance computing and data analysis. NERSC is the principal provider of high performance computing services to Office of Science programs — Magnetic Fusion Energy, High Energy Physics, Nuclear Physics, Basic Energy Sciences, Biological and Environmental Research, and Advanced Scientific Computing Research Key resources within NERSC include Edison, a 2.5 Petflop supercomputer, Hopper, a 1.2 Petaflop supercomputer, a number of mid-scale computational and data analyis clusters and large-scale storage infrastructure.
ESnet	Eli Dart	ESnet provides the high-bandwidth, reliable connections that link scientists at national laboratories, universities and other research institutions, enabling them to collaborate on some of the world's most important scientific challenges including energy, climate science, and the origins of the universe. Funded by the DOE Office of Science, and managed and operated by the ESnet team at Lawrence Berkeley National Laboratory, ESnet provides scientists with access to unique DOE research facilities and computing resources.

Table 4. BER CESD operational and production data portals.

Project	Portal URL	Contact	Description
ARM	http://www.arm.gov/data/	Giri Palanisamy	Includes data quality, data management facility, reprocessing
BER Clearinghouse	http://berdata.ornl.gov/cesd/	Giri Palanisamy	Metadata records from ESGF, ARM, CDIAC and NGEE
CDIAC	http://ameriflux.ornl.gov http://cdiac.ornl.gov/CO2_Emission/ http://cdiac3.ornl.gov/waves/underway/ http://cdiac3.ornl.gov/waves/discrete/ http://mercury.ornl.gov/ocean/	Tom Boden	Supports multiple search and data interfaces/portals
ESGF	http://github.com/ESGF/esgf.github.io/wiki/Peer-Node-Status	Dean N. Williams	Federated data archived at 60+ sites worldwide for over 40 data products
NGEE Arctic	http://ngee-arctic.ornl.gov/	Tom Boden/Giri Palanisamy/Ranjeet Devarakonda/Stan Wullschleger	Includes field, experimental, and model data
SPRUCE	http://mnspruce.ornl.gov/	Misha Krassovski/Les Hook/Paul Hanson	Includes field, experimental, and model data

Table 5. Technologies and services needed for CESD's integrated ecosystem of heterogeneous data.

Data archive	Mgmt.	Data Mining	Provenance & Workflow	Ontology	Analysis	Visualization	Centralized	Federation	HPC
ACME	X	X	X	X	X	X		X	X
AmeriFlux	X	X	X	X	X	X		X	
ARM	X	X	X	X	X	X	X		X
CAPT	X	X	X	X	X	X	X	X	X
CASCADE	X				X	X			
CDIAC	X	X	X	X	X	X	X		
CMIP (PCMDI)	X	X	X	X	X	X		X	X
COSIM									
EMSL	X		X		X				
FACE		X			X		X		
ILAMB									
NGEE	X		X	X	X	X	X	X	X

Arctic									
Obs4MIPs and 70 other model intercomparison projects (MIPs)	X	X	X	X	X	X	X	X	X