Towards a Shared ESS Cyberinfrastructure: Vision and First Steps

Report from the ESS Executive Committee Workshop on Data Infrastructure
August 29-30, 2016.
DOE Headquarters, Germantown, MD

1 Introduction and Motivation

A central objective of the Climate and Environmental Science Division (CESD) within the Office of Biological and Environmental Research (BER) is to advance the predictive understanding of Earth's climate and environmental systems. Assessing the fundamental challenges of this goal led the BER Advisory Committee to assert that "the innovation most needed is a framework that allows seamless integration of multiscale observations, experiments, theory, and process understanding into predictive models for knowledge discovery" (BERAC 2013). The report dubs this innovative framework the Virtual Laboratory. Similarly, projects in Environmental System Science (ESS) have begun exploring requirements for this framework by formalizing the iterative cycle of model-driven experimentation and observation, dubbed ModEx (Figure 1). In all areas, and from both perspectives, significant progress has been made toward achieving the overarching goal. However, significant fragmentation across projects and disciplines remains, making the development of a CESD-wide community-driven cyberinfrastructure an urgent and critical need (U.S. DOE. 2015b). In addition, experimental, observational and computational capabilities are driving exponential growth in the amount, variety and complexity of scientific data (Williams et al., 2014). In combination, these factors suggest the need for a data center that would be a foundational part of a community cyberinfrastructure (Figure 5) that would effectively support the data related needs of TES and SBR projects.

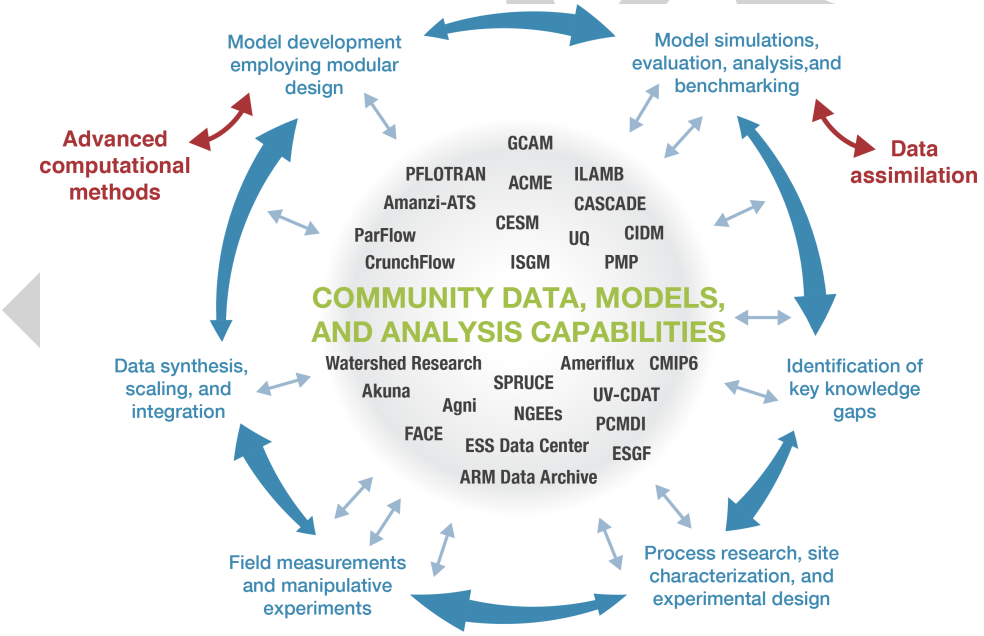

Figure 1. Process studies and field measurements play a critical role in the model–data experimentation (ModEx) enterprise outlined in this diagram. All of these steps both use and produce data, models, and analysis capabilities and tools that can be shared and used by the larger international research community. The goal of the BER Virtual Laboratory is to provide the cyberinfrastructure necessary to automate and support the ModEx enterprise, and the ESS Data Center is an important component of that infrastructure. Figure adopted from the 2016 ILAMB Workshop Report (in press).

The primary objective of this report is to develop and articulate the set of requirements that such a data center must meet to serve the wide range of stakeholders targeted by BER, and the long-term vision of building a virtual laboratory through a community driven and supported cyberinfrastructure.

To develop these requirements, the Executive Committee (EC) of the BER Environmental System Science (ESS) Cyberinfrastructure Working Groups was brought together for a workshop on August 29–30, 2016 at DOE Headquarters in Germantown Maryland. The EC creates and dissolves working groups to address critical needs in the design and development of a community-based cyberinfrastructure. The current working groups include Data Management, Workflows for Model–Data Integration, and Software Engineering and Interoperability. The EC has 10 regular members that span a wide range of ESS projects, and expertise that touches all aspects of the ModEx cycle. This scientific breadth of the EC provides an important holistic connection to the data and workflow dependencies in this iterative cycle, and makes it well suited to developing data center requirements.

The organization of this report builds on the important connection between the evolving ModEx iterative cycle (Figure 1) and the Virtual Laboratory vision to identify requirements for the community-driven cyberinfrastructure, and particularly, the data centers. First, in Section 2, complementary views of this cyberinfrastructure that were developed by different groups and projects within the broader CESD community are highlighted. Section 3 describes a phased approach to implement this collective vision, and identifies the critical need for a high-level design based on modular services. The subsections that follow build from the near-term goals and requirements of the new data center (0–2 years), through a period of enhanced community tool and workflow development for data-model integration and analytics (2–5 years), to predictive understanding enabled by the Virtual Laboratory (5–10 years). For each phase key requirements are identified that influence design choices made in earlier phases. The description of the data center requirements presented here reflects this important connection to the long-term vision.

2 ESS Community Cyberinfrastructure and Virtual Laboratory

The long-term vision of the Virtual Laboratory, and the community-driven cyberinfrastructure that supports it, continues to evolve in concert with the needs of the Climate and Environmental Systems scientists striving to develop a predictive understanding of the Earth System. Within the ESS community the conceptual view of the ModEx iterative cycle (Figure 1) has proven to be a constructive tool for framing this complex multifaceted challenge. In particular, it highlights not only the outer iterative cycle of the primary facets (e.g., Data Synthesis and Model Development), but also the natural feedbacks and cycles that exist within and between the facets that are critical to supporting the modern scientific method. Thus, ModEx provides a useful perspective on the tools and capabilities that have developed across the projects and programs within CESD, as well as the need and opportunity to integrate these tools and capabilities in a community-driven cyberinfrastructure.

Within the Climate community a key driver has been the Coupled Model Intercomparison Project (CMIP). This project has established a worldwide standard for studying general circulation model output, and drives specific complex data management and service requirements. Scientists from BER, in collaboration with agencies around the world, are playing a leading role in evolving a CESD integrated data ecosystem (Williams et al. 2014, 2016) that addresses these requirements. The current vision of this ecosystem is shown in Figure 2, and highlights the leveraging of existing tools and new capabilities in this integrated and federated system. Moreover, it represents an application driven perspective of the broader Virtual Laboratory vision.

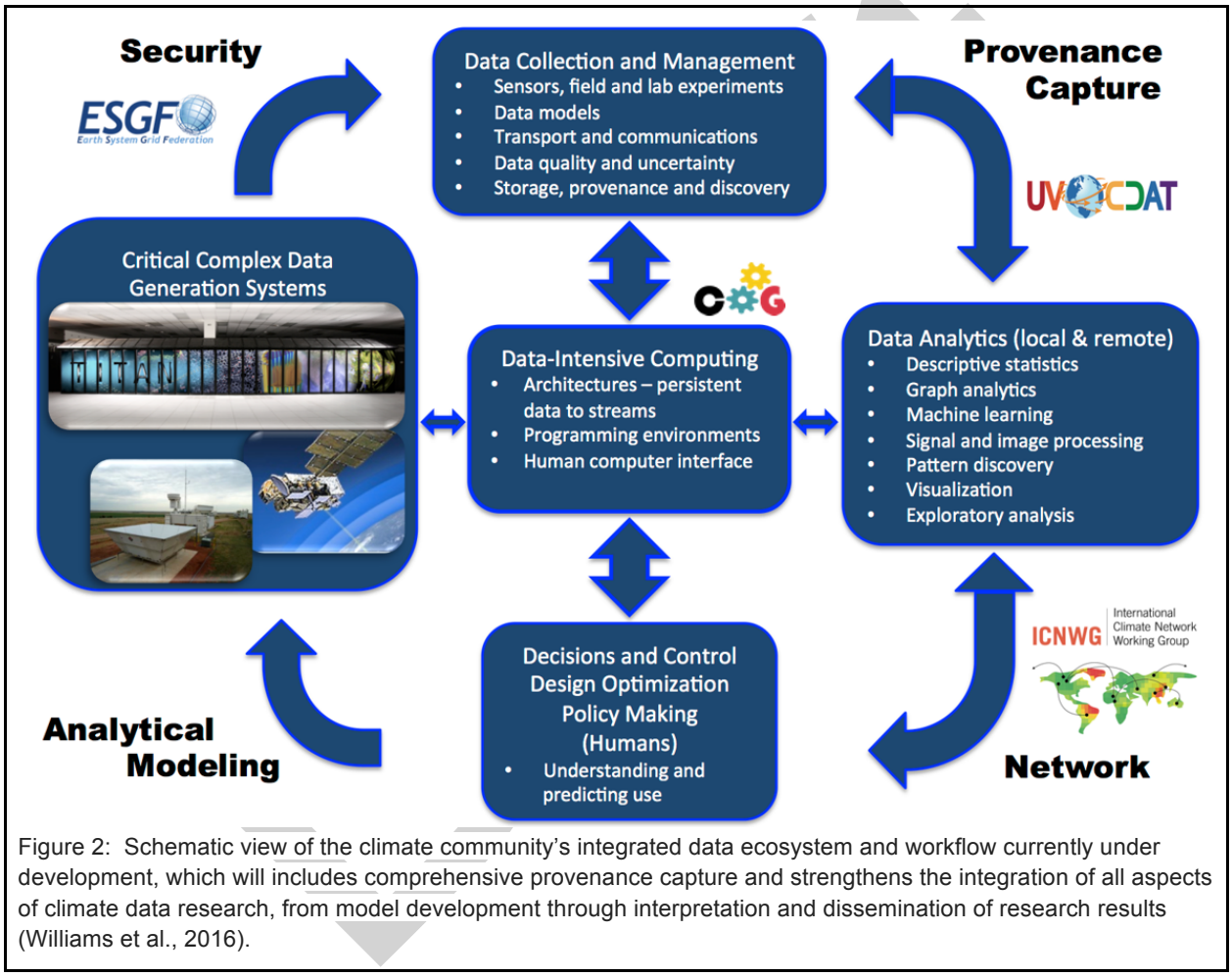

Figure 2: Schematic view of the climate community's integrated data ecosystem and workflow currently under development, which will includes comprehensive provenance capture and strengthens the integration of all aspects of climate data research, from model development through interpretation and dissemination of research results (Williams et al., 2016).

Until recently, much of the ESS science driving data management and service requirements had been focused on the development of process-level understanding at relatively fine scales. Data sets were primarily associated directly with projects, were very heterogeneous, and in many cases did not tie directly to model development. However, within the last three years, ESS has begun to address process-level understanding over a wider range of scales that extend from the bedrock to the planetary boundary layer (which is also known as the atmospheric boundary layer). This focus and its implications for data management, model-data integration workflows,

and software engineering are evident in the workshop report that outlines a phased approach to developing a Virtual Ecosystem capability (U.S. DOE 2015a). To address these common needs across the current ESS projects, a phased development of a community-driven cyberinfrastructure is envisioned that will leverage a dynamic and coordinated set of working groups with expertise in model development, data management, software engineering, and workflows for model-data integration (Figure 3). A new ESS Data Center should provide critical foundational support for this broader community-driven cyberinfrastructure and ModEx approach. In addition, the ESS Data Center should have a well-articulated role within the integrated cyberinfrastructure shown in Figure 3 that supports the broader CESD Data Ecosystem highlighted above (Figure 2).

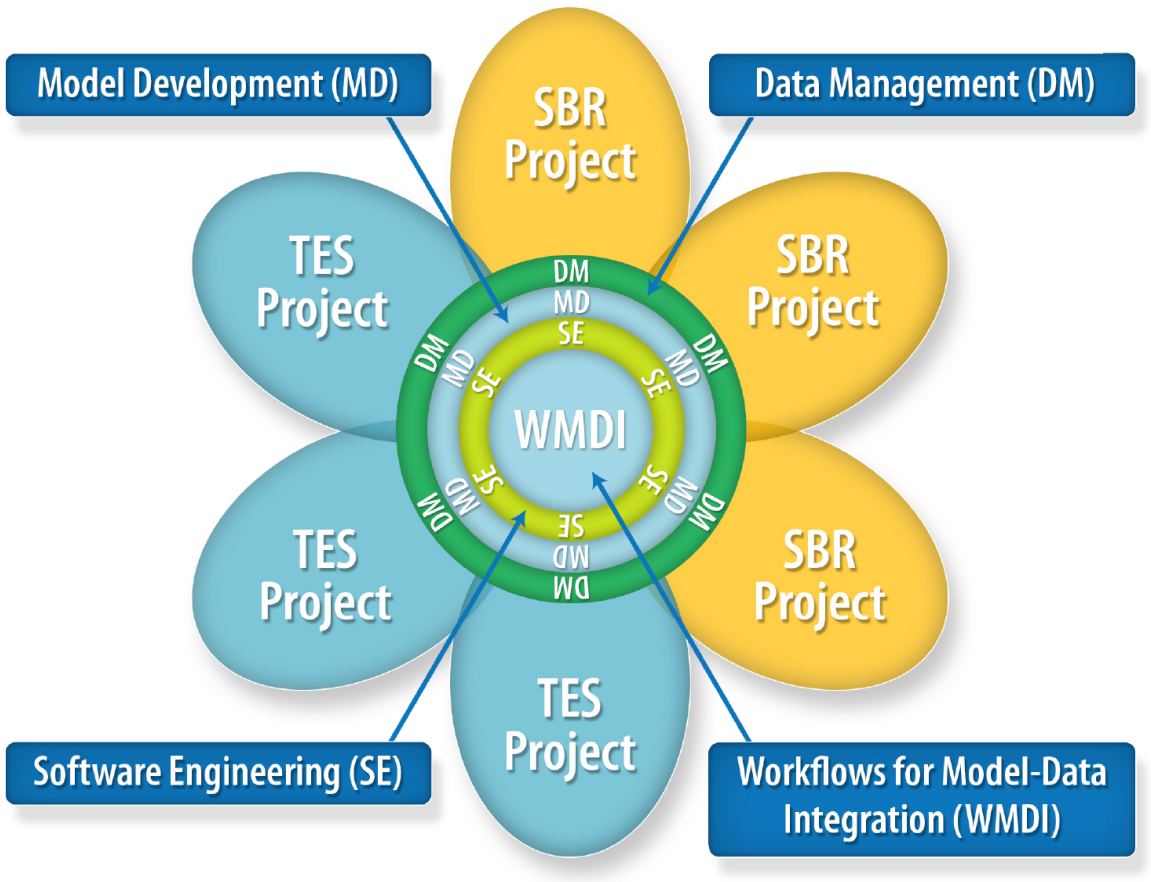

Figure 3: The community-driven cyberinfrastructure for BER Environmental System Science (ESS) will be developed in phases by a dynamic and coordinated set of working groups with expertise in model development, data management, software engineering, and workflows for model-data integration. The integration of these working groups across TES and SBR projects is shown, and the further integration of this cyberinfrastructure is envisioned with development in the broader CESD community and in external agencies (U.S. DOE. 2015b).

Thus, it is apparent that although different science drivers may shift a project's perspective on priorities for the Virtual Laboratory capabilities, a coordinated phased approach driven by the community is needed to ultimately realize its complete vision and impact. Specifically, the Virtual

Laboratory should integrate disciplines and scientists, field sites and field data, experimental and laboratory data, temporal and spatial scales and conceptual and numerical models so that scientists can effectively develop a predictive understanding of natural, managed, and disturbed, terrestrial, aquatic, and subsurface systems across multiple scales and processes. While the actual implementation of this virtual laboratory is likely to be different from any of the representations shown above, the underlying principles will remain the same. The ESS Community Cyberinfrastructure will be an integral component of this virtual laboratory.

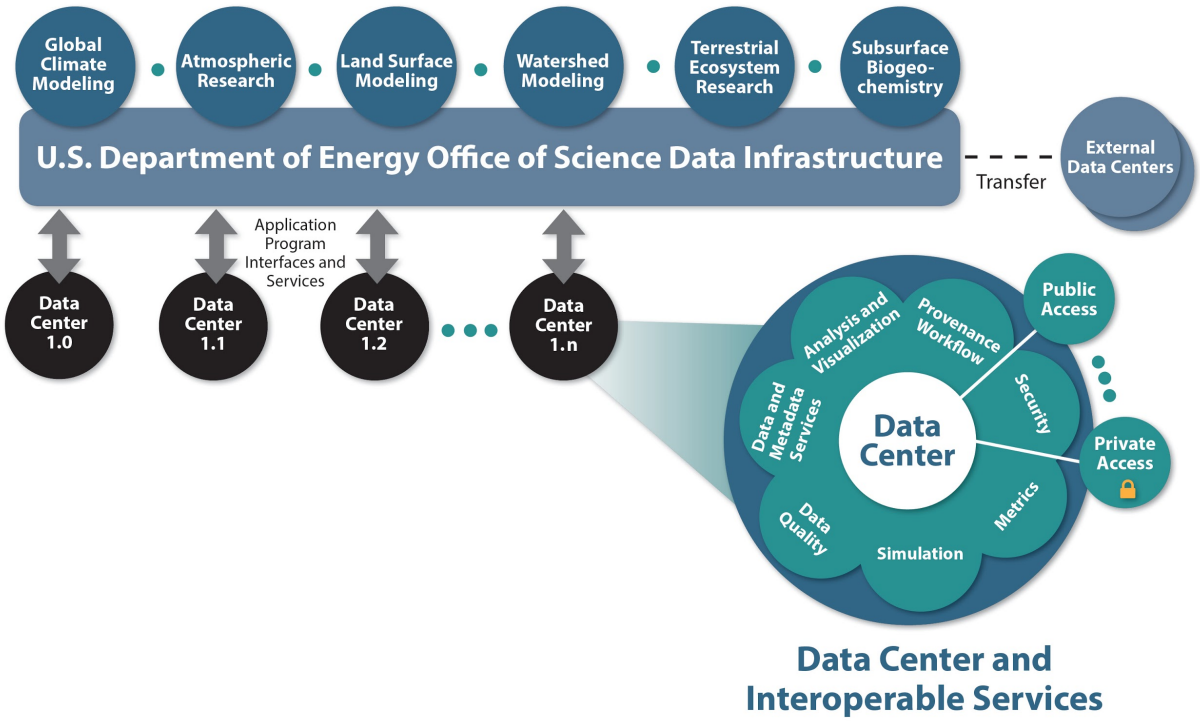

Figure 4. A schematic of an integrated cyberinfrastructure leveraging core DOE Office of Science resources to support critical aspects of the ModEx approach to scientific discovery and predictive understanding (U.S. DOE. 2016).

3 Vision for ESS Community Cyberinfrastructure

The ESS community Cyberinfrastructure should enable world-class science by providing capabilities for data ingestion, management and curation, data analysis and visualization, coupled modeling and data publication. These capabilities should exist in a collaborative research environment that allows for data and result sharing.

In our vision these capabilities would be provided as modular services with well-defined application programming interfaces (APIs). These services can have multiple origins, with some developed by CESD-funded university investigators, while others could be developed by scientists in DOE CESD funded projects (e.g., SFAs, NGEEs). Still other services could be developed under funding from other federal agencies or services developed by commercial

entities (e.g., Google Earth Engine). These services would be implemented within three high-level components (Figure 5) of the ESS cyberinfrastructure, which could be implemented in phases over a period of 10 years. The first phase consists of an ESS Data Center (0–2 years) that would provide capabilities for data archiving, publication, discovery and retrieval, and user support. The second phase (2-5 years) adds a series of community tools that can provide data analysis and visualization capabilities, while continuing to evolve the data center. Finally, in the third phase (5–10 years) virtual laboratory capabilities that facilitate data-model integration and enable knowledge discovery would be implemented, while the data center and the community tools continue to mature.

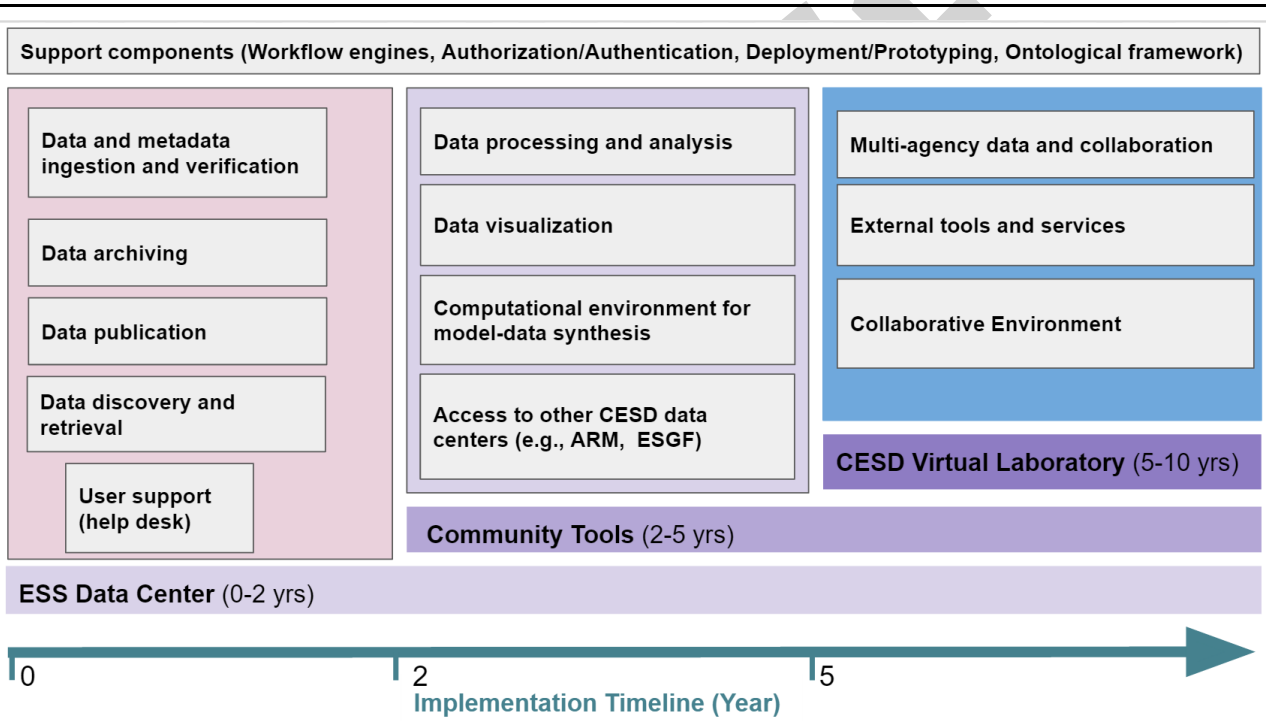

Figure 5. High-level components provided by the ESS Cyberinfrastructure (10 year vision). Each component could consist of modular services that would build upon each other.

As shown above (Figure 5), implementation of these components (and the individual services) would be done in a stepwise fashion. Community involvement from scientists from many different disciplines and research groups will be the key to the successful implementation of these services and adoption by the community. It is clear that while one can define **what** such services should do, any attempt to prescribe an implementation years in advance is futile due to rapid software developments.

However, we can define some basic guiding principles for components and services:

- ⇒ All services should have a clear API.
- ⇒ All services and subservices should (to the extent possible) be modular and interoperable.

⇒ The overall architecture should be open and allow rapid integration of components contributed by the community (CESD funded scientists, general DOE community and scientists not directly funded by DOE).

⇒ All directly funded services which are meant to be implemented in the system in Figure 5 should be developed using an open source approach from the beginning, with rapid development of a Minimum Viable Product as a high level objective. All such services should be designed from the beginning to allow for functionality testing and verification by outside groups.

⇒ Input and feedback from the user community should drive the design of the archive and continual improvement of the data management capabilities. This communication should be facilitated through multiple channels (e.g., publishing a product's roadmap, mailing lists, blogs, and workshops).

⇒ All services should to the extent possible leverage existing efforts and capabilities, and should include as part of the initial roadmap a brief assessment of existing efforts so as to avoid duplication.

3.1 An ESS Data Center Providing Data Management and Services (0–2 years)

The ESS Data Center should provide core capabilities to the ESS community for data and metadata ingestion, curation, archiving, long-term preservation, and publication. These capabilities will form the foundation of the ESS community cyberinfrastructure and should be designed such that they will enable and support the development of community tools and the long-term vision of the BER Virtual Laboratory.

The ESS data center can be a single or federated entity working together to provide the envisioned data management and services. The data center architecture should enable and encourage partnership with existing ESS projects and with the broader ESS community by leveraging project capabilities. To enable these partnerships, the ESS Data center should be guided by a publically available strategic plan that covers the scope of operations through time, that is periodically updated, and that describes the long-term vision for the centers capabilities. Governance of the center should be clearly defined and include roles and responsibilities for the center, ESS projects, and ESS advisory groups.

Key roles and duties for the ESS Data Center that should be included in this strategic plan are 1) Data ingest, archive, and preservation capabilities, 2) Publication tools, 3) Data discovery and retrieval capabilities, 4) Help desk provisions, and 5) a description of data center operations. In addition, this plan should—where possible—layout how the capabilities will support community tool and virtual laboratory efforts (Figure 1).

3.2 Community Tools for Data–Model Integration and Analytics (2–5 years)

The second phase of community cyberinfrastructure development should focus on the creation of community tools and capabilities (Figure 5). New capabilities of an ESS Data Center developed in this phase should expand upon the data ingest and archiving services, as well as

the initial data discovery and retrieval capabilities developed in the first two years (Section 3.1). These capabilities should be consistent with the DOE Virtual Laboratory vision and provide evolutionary functionality toward facilitating predictive understanding of climate and environmental systems as described below (the 5–10 year goals). The focus should be on developing tools and services for the research community, applying new technologies that provide novel modes of data sharing, directly supporting resources for small-to-medium scale (0–50 TB) data analytics and visualization, and integrating services with DOE-ASCR resources in the NERSC, ALCF, and OLCF computing environments for large scale computing and (>50 TB) analytics. Improved support for geospatial data analysis and query, particularly for aerial- and satellite-based remote sensing data, will be increasingly important as these large data streams are fused with in situ and heterogeneous data from ESS field activities. Geospatial query capability is also necessary to automate model setup as well as model–data comparison. In particular, new methods for data delivery may involve remote (server-side) spatial and temporal subsetting and further processing of data, units and format conversion, and support for various network transport protocols to accommodate a plethora of use cases, including on-line data assimilation, model parameter estimation, and model evaluation and benchmarking. All of these processing steps should be automated and robustly integrated, and they must support computational environments from small clusters to the leadership class facilities (LCFs) with one-time authentication (single sign-on). Further integration with model output archives and delivery systems, such as the Earth System Grid Federation (ESGF), is expected to be important for combining observations and model results to answer various research questions and support model benchmarking. Sharing and co-developing data format specifications, APIs, and transport and delivery methods between model and measurement data systems may facilitate a convergence of these systems in the future and offer researchers a unified Federated platform for data processing, modeling, and analytics. One possible solution for delivering these needed services is a BER Science Cloud platform that offers codes, easy access to data holdings, automatically retrieved collections of remote sensing data. Such a platform could be at the heart of what might grow into a larger facility in future years.

During this second phase the role (and potentially the implementation) of an ESS Data Center is expected to evolve in response to these needs. In this phase an ESS Data Center should offer services and resources (hardware and software) to support synthesis, data analytics, and meta-analysis for groups of BER researchers, including collaborative working groups modeled on those in the National Center for Ecological Analysis and Synthesis (NCEAS) that may be sponsored by other BER Programs. Possibly later a NCEAS-style meeting facility with analysis and visualization resources accompanied by high-speed access to data for on-site analytics and knowledge discovery could be developed. Such NCEAS-style working groups would bring together scientists from the wider international research community to work with BER scientists. To be truly successful, an ESS Data Center would need to work across DOE facilities and consider how best to also integrate with external data centers (e.g., NASA DAAC, NOAA NCEI, etc.) to support BER research needs.

3.3 BER Virtual Laboratory for Improving Predictive Modeling (5-10 years)

An overarching goal of the CESD is to improve our predictive understanding of the coupled Earth system (i.e., atmosphere, land, ocean, sea ice, land ice) under various, atmospheric emission and exchange, land use, and other perturbation scenarios over the next century. Currently, uncertainty in Earth system models (ESMs) is dominated by uncertainty in initial conditions, model structures, model parameters, spatial scaling, and numerical methods. Enormous community efforts are underway to improve the ESM predictability and reduce uncertainty through tight coupling of field observations, laboratory experiments, and model development and benchmarking as shown in Figure 2. In 2015, DOE convened a Working Group on Virtual Data Integration to lay the groundwork for a federated BER Virtual Laboratory and CESD data infrastructure (Williams et al., 2014, 2016). Through extensive analogies with a handful of use cases, the resulting report identified critical needs for publishing and archiving data; comparing diverse data types; supporting data access and usage; obtaining observational, experimental, computational, and storage resources; among others. The Working Group offered suggestions for addressing these challenges and established a vision for an integrated and Federated Virtual Laboratory to automate data collection, processing, archiving, modeling, analysis, and publishing.

The ESS community within CESD provides the multi-scale mechanistic understanding of natural, disturbed, and managed terrestrial ecosystems extending from the bedrock to the top of the planetary boundary layer that underpins improvements in modeling climate impacts and feedbacks on watershed and ecosystem function, water resource management, and biogeochemical cycles. The design of an ESS community cyberinfrastructure (U.S. DOE, 2015b) should thus provide a collaborative environment to enable leveraging and sharing of data and information within the CESD community and beyond (multi-agency collaboration). A variety of technologies and services, including use of modular APIs for data exchange, are necessary to benefit from the wide range of temporally and spatially heterogeneous data and much more complex model structures being planned over the next decade.

We envision the model parameter and structural estimation efforts highlighted above in the 2–5 year vision to continue in the third phase of the ESS Data Center for years 5–10. However, additional analyses should support ensemble simulations of ESMs (e.g., through the Advanced Climate Modeling for Energy (ACME) and the Community Earth System Model (CESM) frameworks), across multiple DOE computing centers, and tightly coupled with a variety of observational datasets (from ESS and multi-agency data centers) for model testing and analyses. High performance computing systems, flexible workflow tools, modular process models, machine learning and scaling algorithms, and similar capabilities are required to robustly characterize sensitivity and uncertainty in predictions (e.g., from posterior parameter distributions and model structural uncertainty). The focus for a third phase of an ESS Data Center should be on enhancing services, tools, and integration with this cyberinfrastructure to realize the vision of a BER Virtual Laboratory, which would facilitate site, watershed, regional, and global predictions with quantified uncertainties. The resulting cyber infrastructure should be designed to facilitate improvements in mechanistic representations in models.

4 Suggested Phase 1 ESS Data Center Requirements

The shifting needs of Climate and Environmental System scientists are driving the demand for a new community-driven cyberinfrastructure that can support the long-term vision of the Virtual Laboratory. The ModEx conceptualization (Figure 1) of this long-term vision provides a framework to help realize a phased development strategy in which the ESS Data Center plays a foundational role. The following subsections highlight the requirements for this Data Center.

4.1 Data ingest, archiving, and preservation capability

The data center should provide metadata development tools for ESS projects, and provide guidance to ensure the use and application of community standards and conventions for data formats. Standards and conventions should be jointly established between ESS community and data center staff. The data center in collaboration with the ESS community should enforce the inclusion of a set of high-level information and format requirements in all metadata, and also provide a more comprehensive listing of best-practice variables to be included in the description of observation and experimental data. An automated data entry system for metadata could check for the inclusion of key variables and formats.

The data center should develop and provide data archiving services and capabilities which would support both project and broader ESS community needs and allow for long term access to preserved data post project completion. Such archiving and preservation services should be user friendly and support versioning of data products.

A key near-term goal of the ESS Data center is to provide continuity for a subset of the Carbon Dioxide Information Analysis Center (CDIAC) capabilities, data holdings, and tools. DOE program managers will define the specific components of CDIAC that should be transferred to the new ESS Data center.

4.2 Ontological tools

The data center should assist in developing ontological tools that allow ESS scientists to (where feasible) map project specific names to standard names and units. This effort should be driven by the anticipated need for model and data coupling.

4.3 Publication tools

The data center should provide a system for ESS scientists to acquire persistent data identifiers (e.g. currently represented by the digital object identifier, DOI) for the data sets that they generate. The data center should also establish a standard for the development of data sets intended to accompany publications, and provide recommendations for the DOI citations, expected acknowledgements, and dataset contributor notifications.

4.4 Data discovery and retrieval

The data center should provide online capabilities to search across central and project developed and maintained ESS data and metadata catalogues. Capabilities for searching and retrieval of data should allow both human and programmatic access. Data retrieval in different well established and community accepted formats should be supported.

4.5 Help desk

The data center should provide and support staff to operate a help desk to enable ESS researchers to use and understand the data archive and retrieval systems. Help desk personnel should be able to provide instructions, and explain policies, procedures, and standards expectations to the center users.

4.6 Data center operations

The data center should calculate and provide data and tool usage metrics to track the engagement of the data center with the research community. The data center should, where possible, support single sign on for ESS researchers. The data center should be highly available through the Internet with a minimum of down time. Plans for a new data center should explain how this will be made possible. Appropriate security and authentication mechanisms should be established to enable usage tracking and enforce usage rules. Where data use policies place restrictions on the dissemination of data (this should be minimized) the center should be responsible for the enforcement of such policies. Data center personnel should be active participants in research meetings and workshops to gather community input and perform community outreach and training.

4.7 Community software tools

The data center should be an active participant in the development and support of the ESS community cyberinfrastructure. In the near term, all tools that are developed and supported by the ESS data center should be well-documented open-source software with associated deployment and testing procedures. Plans to support and distribute these open-source tools should be described in detail and vetted by the broader community.

In the later years (~2–5 years), as community tools and infrastructure are being actively developed, the data center should be prepared to support this development in two ways. First, it should actively participate in requirements gathering for tools developed by the community (outside the data center), as well as support their installation and use at the data center. Second, in cases where the data center has a critical need and expertise, the data center should consider developing and supporting new capabilities to enhance the emerging ESS community cyberinfrastructure.

5 References

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, Washington D.C., (genomicscience.energy.gov/program/beracvirtuallab.shtml).

U.S. DOE. 2016. *Working Group on Virtual Data Integration* [pdf]*: A Report from the August 13-14, 2015, Workshop.* DOE/SC-0180. U.S. Department of Energy, Office of Science, Washington, D.C.

U.S. DOE. 2015a. *Building Virtual Ecosystems: Computational Challenges for Mechanistic Modeling of Terrestrial Environments: Workshop Report*, DOE/SC-0171. U.S. Department of Energy Office of Science, Washington, D.C. (http://doesbr.org/BuildingVirtualEcosystems/).

U.S. DOE. 2015b. *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, Washington, D.C. (http://doesbr.org/ESS-WorkingGroups/).

U.S. DOE. 2014a. *Extreme-Scale Scientific Application Software Productivity* [pdf], report prepared for the Office of Advanced Scientific Computing Research (ASCR) Workshop on *Software Productivity for Extreme-Scale Science*, Hilton Rockville, Rockville, MD, January 13–14, 2014.

U.S. DOE. 2014b. *Data-Model Needs for Belowground Ecology: A Summary Report from the Terrestrial Ecosystem Science (TES) Mini-Workshop*. U.S. Department of Energy Office of Science (http://tes.science.energy.gov/workshops/).

U.S. DOE. 2002. "Policy Guidance – OSS License Release of Software Developed with ASC and OASCR Funding," Memo from the U.S. Department of Energy Offices of Advanced Simulation and Computing and Advanced Scientific Computing Research (http://science.energy.gov/~/media/ascr/pdf/research/docs/Doe_lab_developed_software_policy.pdf).

Williams D.N. et al. 2016. "A Global Repository for Planet-Sized Experiments and Observations", *Bulletin of the American Meteorological Society*, June, pp. 803-816, DOI:10.1175/BAMS-D-15-00132.1.

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

Appendices

Appendix A: Workshop Agenda

ESS Executive Committee Data Infrastructure Requirements to Support Current and Future Science Projects

DOE Headquarters, Germantown, MD
August 29–30, 2016

Monday August 29, 2016	
Time	**Topic**
9:00 am - 9.15 am	Welcome and introduction (Gary Geernaert) Workshop charge (Jay Hnilo)
9:15 am – 9:30 am	Identifying ESS computational and data environment (Paul Bayer, David Lesmes, Dan Stover and Jay Hnilo)
9:30 am– 10:00 am	**Science Drivers** Discussion Lead (David Lesmes) • Example use case requirements (Jay Hnilo) 10 mins • Define what are the key things that are difficult to do today and are impeding scientific progress or productivity • Science case discussion (20 mins) (list future science drivers; what are their limitations)
10:00 am - 10:30 am	**Break**
Directed Roundtable Discussions	
10.30 am – 11:15 pm	**Data Services to Support Science Requirements** Discussion Lead: David Moulton Scribe: Roelof Versteeg Questions: • What are the key challenges that ESS scientists encounter? • What data services would address the identified challenges? What exists already today? What do we still need? What are the key characteristics that these services need to have to be successful (i.e. integrated, easy to customize etc.)? • What are the key impediments (on the data provider / 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?

11:15 pm - 12:30 pm	**Required Data Center and Interoperable Services for ESS** Discussion Lead: Roelof Versteeg Scribe: Forrest Hoffman Discuss and rank data center services required in an integrated infrastructure A preliminary list (open to modification during the meeting) of these services includes 1. Data and metadata ingestion and verification services 2. Data discovery and retrieval services. 3. Data publication services 4. Data visualization services 5. Data processing and analysis services 6. Modeling services 7. Supporting services. This would include components which support all other services, and would include e.g. a. workflow services b. authorization/authentication/access services c. deployment/testing services
12:30 pm - 1:30 pm	**Lunch**
Directed Roundtable Discussions (Continued)	
1:30 pm - 2:15pm	**Inventory of existing ESS and CESD data tools and services, benchmark of tools for potential reuse** Discussion Lead: Deb Agarwal Scribe: Paul Hanson Questions: ● What tools have been identified during the previous discussions that should be made more widely accessible to the ESS community (Libraries)? ● How should we evaluate and decide on tools to adopt? ● 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 tool developer and community? ● How do we want to assess the maturity and capability of tools (e.g. benchmarks or crowdsourcing)? ● What are the interface standards available or that need to be developed to enable a common tools and services ecosystem?
2:15pm - 3:15pm	**Advanced Computational Environments and Data Analytics** Discussion Lead: Forrest Hoffman Scribe: Xingyuan Chen 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: ● Define a scalable compute resource (clusters and HPCs) for ESS 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 Identify key gaps, identify benefitting communities, and prioritize
3:15 pm- 3:45 pm	**Break**
3:45 pm- 4:30 pm	**Data Services and Monitoring** Discussion Lead: Eric Pierce Scribe: Bill Riley Questions: ● What data services and monitoring are required to support ESS research community needs? ● How can these data services and monitoring support the research community? ○ **Services:** Ready and easy access to data products, monitor data downloads (statistics), and track data products created as a result of downloads (statistics). ○ **Services:** Track users that access data products (statistics). ○ **Services:** Reliable access to data archive system. ○ **Services:** Ready and easy access to sufficient documentation (metadata, but also experimental details/data collection details – more than simplified metadata). ○ **Tools:** Enable easy discovery and data query. ○ **Tools:** Enable access to post-processed data (version control) and that enable access to real-time data being logged (pie in sky). Also archive all versions of data products. ○ **Tools:** Provide alerts to complementary and future data that is similar in nature (Home Depot example where the products are advertised based on interest). This type of profile approach can enable scientific discovery.
4:30 pm - 5:15 pm	**Participation with broad/multi-agency data initiatives** Discussion Lead: Bill Riley Scribe: Stan Wullschleger Topics: ● Standards and services that needs to be adopted within the compute environment that will allow ESS and CESD to participate in multi-agency data initiatives such as EarthCube, USGEO etc. ● Data sharing with NASA DAACs, NOAA, and other agencies
5:15 pm - 5:30 pm	**Wrap up Discussion and Planning** Discussion Lead (David Moulton) ● Identify critical topics that need further discussion and identify leads ● Review current requirements and layout for report ● Finalize schedule for Tuesday
	Dinner

Tuesday August 29, 2016	
Time	**Topic**
8:30 am - 8:45 am	**Recap/Status and Schedule** Discussion Lead (David Moulton)
8:45 am - 9:15 am	**Revisit Topic 1 (TBD)**
9:15 am - 9:45 am	**Revisit Topic 2 (TBD)**
10:00 am - 10:30 am	**Break**
10:30 am - 12:15 am	**Action Items and Writing** • Finalize layout and writing assignments for the report • Set schedule for EC meetings and writing deadlines for report • Begin writing
11:45 am - 12:15 pm	
12:15 pm - 1:15 pm	**Lunch/Adjourn**
1:15 pm- 3:00 pm	Additional writing time for those that can stay

Appendix B: Workshop Organizers and Participants

Deb Agarwal	Lawrence Berkeley National Laboratory
Paul Bayer	DOE-BER-CESD
Xingyuan Chen	Pacific Northwest National Laboratory
Jared deForest	DOE-BER-CESD
Gary Geernaert	DOE-BER-CESD
Paul Hanson*	Oak Ridge National Laboratory
Jay Hnilo (co-organizer)	DOE-BER-CESD
Forrest M. Hoffman	Oak Ridge National Laboratory
David Lesmes	DOE-BER-CESD
Sally MacFarland	DOE-BER-CESD
David Moulton (co-organizer)	Los Alamos National Laboratory
Eric Pierce	Oak Ridge National Laboratory
Bill Riley	Lawrence Berkeley National Laboratory
Dan Stover	DOE-BER-CESD
Roelof Versteeg	Subsurface Insights, LLC
Dean Williams*	Lawrence Livermore National Laboratory
Stan Wullschleger*	Oak Ridge National Laboratory

**Attended workshop by teleconference*

Appendix C: ESS Executive Committee Members

Members are listed alphabetically.

Name	Institution	Expertise	BER Projects
Deb Agarwal	Lawrence Berkeley National Laboratory	Data management, meta data, tools and APIs, and provenance	Ameriflux, NGEE Tropics, LBNL SBR SFA
Xingyuan Chen	Pacific Northwest National Laboratory	Data-model integration, multiscale modeling, SA, UQ	PNNL SBR SFA
Paul Hanson	Oak Ridge National Laboratory	Carbon cycle and experimental manipulations	ORNL TES SFA (SPRUCE)
Forrest M. Hoffman	Oak Ridge National Laboratory	Earth system modeling, global biogeochemistry, model benchmarking	Biogeochemistry–Climate Feedbacks SFA, ACME, NGEE Arctic, NGEE Tropics
David Moulton*	Los Alamos National Laboratory	Modeling, algorithms, computational science, software engineering	Interoperable Design of Extreme-scale Application Software (IDEAS)
Eric Pierce	Oak Ridge National Laboratory	Geochemistry, solid-fluid interactions, experiments and data curation	ORNL SBR SFA
Margaret Torn	Lawrence Berkeley National Laboratory	Ecology, biogeochemistry, carbon cycling, climate	Ameriflux
Roelof Versteeg	Subsurface Insights	Data management, geophysical monitoring	LBNL SBR SFA
Dean N. Williams	Lawrence Livermore National Laboratory	Data management, tools, federation, networks, provenance, visualization	ACME, ESGF, UV-CDAT
Stan Wullschleger	Oak Ridge National Laboratory	Field environmental studies	NGEE-Arctic

Appendix D: List of Acronyms and Abbreviations

ACME	Advanced Climate Modeling for Energy
ALCF	Argonne Leadership Computing Facility
API	Application programming interface
BER	Office of Biological and Environmental Research
BERAC	Biological and Environmental Research Advisory Committee
CDIAC	Carbon Dioxide Information Analysis Center
CESD	Climate and Environmental Sciences Division
DAAC	Distributed Active Archive Center
DOE	U.S. Department of Energy
DOI	Digital object identifier
EC	ESS – Executive Committee
ESM	Earth System Model
ESS	Environmental System Science
LCFs	Leadership Computing Facilities
NASA	National Aeronautics and Space Administration
NCEAS	National Center for Ecological Analysis and Synthesis
NCEI	National Center for Environmental Information
NERSC	National Energy Research Scientific Computing Center
NGEEs	Next Generation Ecosystem Experiment projects
NOAA	National Oceanic and Atmospheric Administration
OLCF	Oak Ridge Leadership Computing Facility
SA	Sensitivity Analysis
SFAs	Laboratory Scientific Focus Areas
UQ	Uncertainty Quantification