LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-TR-654240

Climate Science for a Sustainable Energy Future Test Bed and Data Infrastructure

Final Report

D. N. Williams, I. Foster, K. Kleese-Van Dam, G. Shipman

May 12, 2014

Disclaimer

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

Climate Science for a Sustainable Energy Future
Test Bed and Data Infrastructure
Final Report

Dean N. Williams[1], Ian Foster[2], Galen Shipman[3], Kerstin Kleese-Van Dam[4]

[1] *Lawrence Livermore National Laboratory*
[2] *Argonne National Laboratory*
[3] *Oak Ridge National Laboratory*
[4] *Pacific Northwest National Laboratory*

Abstract

The collaborative Climate Science for a Sustainable Energy Future (CSSEF) project started in July 2011 with the goal of accelerating the development of climate model components (i.e., atmosphere, ocean and sea ice, and land surface) and enhancing their predictive capabilities while incorporating uncertainty quantification (UQ). This effort required accessing and converting observational data sets into specialized model testing and verification data sets and building a model development test bed, where model components and sub-models can be rapidly evaluated. CSSEF's prototype test bed demonstrated, how an integrated testbed could eliminate tedious activities associated with model development and evaluation, by providing the capability to constantly compare model output—where scientists store, acquire, reformat, regrid, and analyze data sets one-by-one—to observational measurements in a controlled test bed.

The CSSEF-DIT team worked closely with climate scientists on the tools for evaluating model components and with user facilities to deploy the prototype test bed. In addition, team members collaborated extensively with national and international institutions and universities that specialize in data-intensive science and exascale computing. The goal was for the new climate-model test bed to incorporate:

- A **calibration platform** where UQ techniques are used to calibrate a model against regional observational data sets; and
- A **validation platform** where simulation quality is quantified against global observational data sets.

The team designed the prototype test bed infrastructure so that it could be easily customized to the specific requirements of each science component area, and demonstrated its capabilities on one of them. The DIT team integrated closely with the modeling activities and UQ methods, and provided interactive access to the (derived) data products to support the goals of CSSEF science community.

Introduction

Under the leadership of Lawrence Livermore National Laboratory (LLNL) and in partnership with Argonne National Laboratory (ANL), Oak Ridge National Laboratory (ORNL), Pacific Northwestern National Laboratory (PNNL), and Sandia National Laboratory (SNL), a nationally federated, distributed data archival and retrieval system was established under the Earth System Grid Federation (ESGF) architecture. In addition, under the Ultra-scale Visualization Climate Data Analysis Tools (UV-CDAT), a framework for diagnostics and the incorporation of UQ was established. For ESGF, UV-CDAT, and other components, the CSSEF team developed the prototype of a new end-to-end prototype user environment to

assist uncertainty quantification through the provision of (a) a test bed for the rapid development and assessment of new model components, (b) relevant data products, and (c) data and computing infrastructure.

The new user environment empowers CSSEF scientists to engage in data exchanges that could ultimately lead to breakthrough climate-science discoveries. This team's effort achieved the following proposal goals:

1. Completed the initial development of the ESGF distributed enterprise system and install the ESGF node software stack at several CSSEF locations (including LLNL, ORNL, ANL, NERSC/LBNL, and SNL);
2. Established the beginnings of a federated CSSEF distributed data archive;
3. Collaborated with cross-cutting CSSEF component teams to produce land, atmosphere, and ocean workflow use cases and requirements (this included designs for UQ efforts in support of sensitivity analysis efforts);
4. Collaborated with cross-cutting CSSEF component teams to produce land, atmosphere, and ocean diagnostics (i.e., AMWG, OMWG, LMWG, etc.)
5. Developed CSSEF interactive and exploratory analysis and visualization; and
6. Build upon DOE open source efforts to produce a workflow and diagnostic prototype test bed.

During the project, the team met four to six times a week to discuss the challenges and needs (shown in **Tables 1 and 2**).

Table 1. *Grand challenges for the complex CSSEF model development test bed.*

Science Need	Solution
Wide range of model runs and workflow types across the project	Define use cases: 1) Model Development; (2) Exploratory Model Runs; and (3) Production Model Runs
Capture and record suites of runs and their settings during model development	Automated provenance and archiving
Quickly evaluate coupled model behavior	Diagnostics of the coupled system within one software system

Table 2. *Major challenges addressed in the development of the prototype test bed.*

Challenge	Description
Installation	Software must be adapted to multiple hardware platforms and operating systems located throughout the CSSEF
Heterogeneous Data Sets	The same infrastructure must allow scientists to access and compare data sets from multiple sources, including from observational satellite and instrument sources
Analysis, Diagnostics, and Visualizations	Better model development and evaluation requires new and improved analysis, diagnostics, and visualization techniques
Server-Side and In Situ Computing	Server-side and in situ computation is necessary as the increase in data size and complexity of algorithms lead to data-intensive, compute-intensive challenges for diagnostics, UQ, analysis, model metrics, and

	visualization

The team also discussed the progress of the CSSEF infrastructure components (shown in **Table 3**) and devised a plan to modify and stand up ESGF nodes at five DOE sites (LLNL, ORNL, PNNL, ANL, SNL) and one NASA site (JPL)—the NASA site was used to obtain needed observational data.

Table 3. *CSSEF open source infrastructure components to build the prototype test bed.*

Tools	Description
Akuna	Advanced Simulation Capability for Environmental Management (ASCEM) developed AKUNA framework, led by PNNL (http://akuna.labworks.org/)
Diagnostics	Diagnostics framework, led by LLNL, co-led by ORNL, LANL, produced substantially equivalent AMWG, OMWG and LMWG sub-model diagnostics along with many others including model metrics (http://uv-cdat.org)
EDEN	Exploratory Data analysis Environment, led by ORNL, is a visual analytics tool for exploring multivariate data sets. (http://cda.ornl.gov/projects/eden/)
ESGF	Earth System Grid Federation (ESGF), led by LLNL, is a worldwide federation of climate and computer scientists deploying a distributed multi-petabyte archive for climate science (http://esgf.org)
ESnet	ESnet: DOE Energy Science Network (https://www.es.net/)
Globus	Globus, led by ANL, is a set of hosted services for data transfer, data sharing, and identity/group management (https://www.globus.org/)
Leadership Computing Facilities (LCF) Infrastructure	Includes ANL's and ORNL's LCFs and the National Energy Research Scientific Computing Center (NERSC). For time period, only ORNL's LCF was used.
ProvEn	Provenance Environment, led by PNNL, is the component integration of the workflow and provenance
Ultra-scale Visualization Climate Data Analysis Tools **(UV-CDAT)**	UVCDAT, led by LLNL, provides access to large-scale data analysis and visualization tools for the climate modeling and observational communities (http://uv-cdat.org)

The work plan included setting up dedicated machines exclusively for the CSSEF infrastructure (at ORNL's LCF).

Project Design

The CSSEF team worked closely with the cross-cutting CSSEF model component and computational research themes to identify requirements and design the CSSEF test bed architecture. **Figure 1** is a high-level conceptual view of the CSSEF test bed architecture and workflow with provenance capture. Provenance capture works with all model components

(atmosphere, land, and ocean) and is pervasive throughout the test bed, capturing and saving all user actions. Getting provenance capture to work throughout the test bed is still in progress. The red arrows show the baseline ensemble loop in which model simulations are conducted using a variety of input parameters generated by Metrics and UQ ensemble drivers. At any stage, data can be collected and stored in the ESGF distributed archive. The black arrows show how desktop clients, Web browsers, or scripts gain access to the test bed.

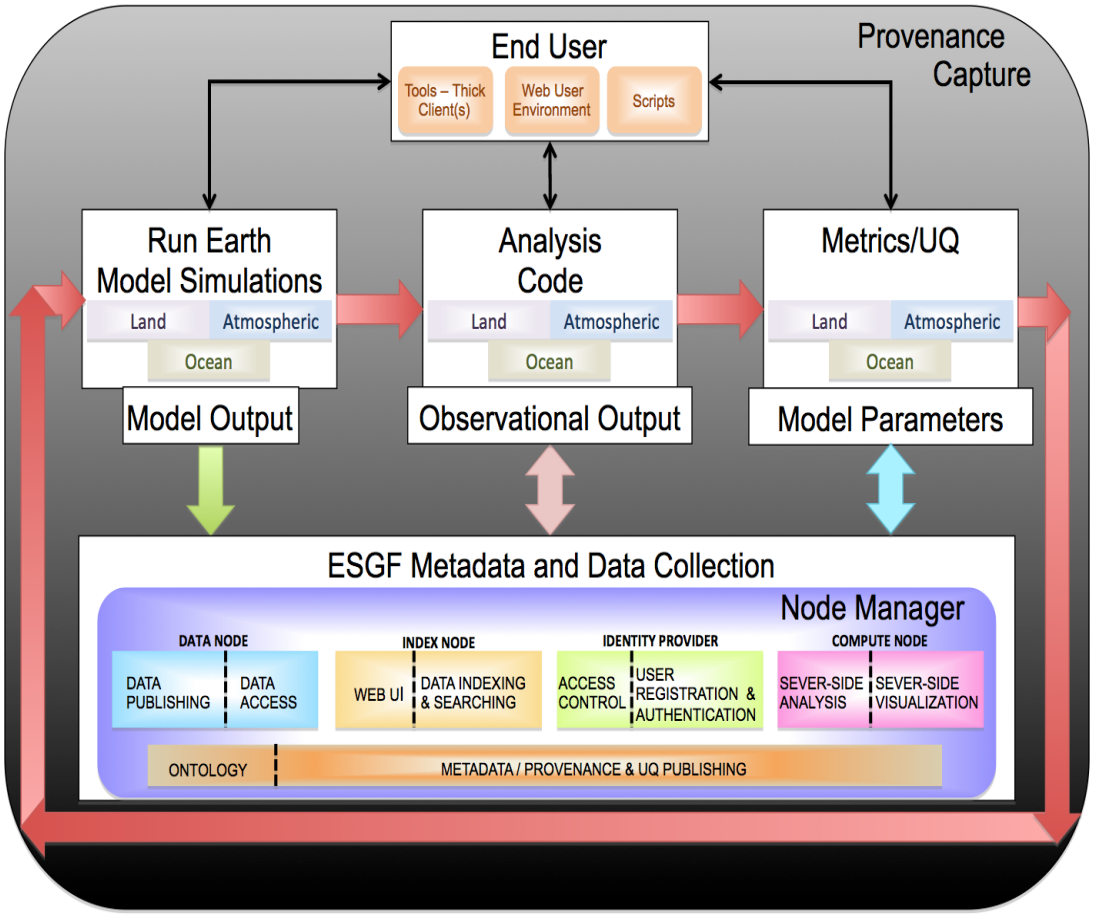

Figure 1: The CSSEF test bed architecture showing the integration of underlying services (see **Table 3**) to facilitate search and run model components and sub-models.

Implementation of the Project Design and Data Summary

To achieve the goals of the CSSEF test bed, the team built upon the efforts of ESGF, UV-CDAT, Globus, ASCEM/AKUNA. These open-source projects, listed in **Table 3**, have growing recognition and use in the research community, and the tools and experience resulting from these DOE-sponsored projects provided the foundation on which the prototype model test bed infrastructure was based. By building upon and integrating these existing technologies, open standards, and community expertise, we built a unique turnkey and flexible framework suitable for accelerating future DOE's Earth system model development that includes integrated workflows and provenance, diagnostics, analysis and visualization, and automated testing and evaluation capabilities.

ESGF Metadata and Data Collection

In the CSSEF community, scientists need to share, access, search, and compute data—primarily using the NetCDF format. Because the community is composed of many large and small institutions dynamically and independently interacting, the ESGF architecture is based on the concept of institutions hosting a dynamic system of nodes that interact on an equal basis and offer a broad range of user and data services depending on how they are set up.

Corresponding to the "ESGF Metadata and Data Collection" in **Figure 1**, an ESGF node can be configured to possess one or more functionalities:

- Data Node: provides services to publish and serve scientific data through a variety of protocols such as HTTP, GridFTP, and OPeNDAP DAP.
- Index Node: provides services to harvest metadata associated with published data and enable data discovery.
- Identity Provider: provides security facilities to register, authenticate, and authorize users.
- Compute Node: provides application services for server-side data reduction, analysis, and visualization of published data.

In this distributed environment, each ESGF node advertises its capabilities (public keys, functionality, service endpoints, etc.) in a registry document, which is continuously propagated to all other nodes through a "gossip" protocol. A gossip protocol is a style of computer-to-computer communication protocol inspired by the form of gossip seen in social networks. Each node is constantly aware of the full state of the federation, and nodes can join and leave the federation dynamically, without impacting the availability of data and services at any other node.

Ontology and Provenance

CSSEF requires a cutting-edge ontological solution that supports diverse, large-scale, and complex computations, analytics, and data. The team relied upon well-established practices and standards found in the semantic Web and scientific communities for an ontology that is a composite of different regions and sub-models centered on the published data sets and processing expected to occur in the CSSEF test bed environment.

CSSEF uses the Provenance Environment (ProvEn) framework for the prototype test bed. Although not fully implemented, ProvEn provides a comprehensive provenance framework with services for the collection and storage of processing provenance. ProvEn correlates processing provenance with knowledge provenance to help scientists browse, query, and infer from their unique scientific perspective. ProvEn will leverage the ontology regions, providing a seamless integration among processing provenance, local metadata, and/or shared community ontologies.

Running the Earth Model Simulation Components in an Integrated User Interface Environment

The new collaborative user environment backed by AKUNA offers a user interface (UI) that enables users to create, edit, and organize their simulation runs. AKUNA orchestrates the management and initiation of new workflow runs and provides an integration point for provenance metadata capture. Existing analysis tools, such as UV-CDAT and other exploratory analysis tools, are also accessible from the UI. Integrating ESGF allows for automated data set retrieval and the storage of selected results generated by the UV-CDAT analysis tools or simulation runs. Although the test bed is designed to work with ESGF, the test bed will also support users who may need to work off-line from the shared ESGF space.

Any provenance metadata and simulation data sets created by users while off-line may be pushed back to ESGF at their convenience. These capabilities will support the CSSEF goal of making it easier to rapidly prototype and evaluate model component changes.

Data Analysis and Uncertainty Quantification and Metrics

The architecture diagram in **Figure 1** shows the workflow of the analysis-code component operating on data from model simulation runs or data archived in ESGF, potentially including additional model simulations and observational data. At any time, analysis may be captured for display or stored in the ESGF archive for later analysis or future study. The analysis-component also provides framework for UQ and model metrics. Analyses are used by UQ to reduce parameter dimensions for further narrowing of viable parameter ranges. In addition, routine performance metrics are captured as scalar quantities to ascertain a model's improvements, the rate of improvement, and the reality of the predictive model compared to observations. These are but two examples of the usefulness of the analysis code.

To accommodate the wide range of existing and potentially new analysis and diagnostics codes for model runs and observation data, the CSSEF test bed analysis code relies on the UV-CDAT framework. The UV-CDAT analysis code is integrated into the ESGF Compute Node, where the data is co-located and can access the Index Node to retrieve data for local use. Within this ESGF–UV-CDAT environment, the CSSEF community members have access to several capabilities:

- Parallel visualization and analysis tools (exploiting parallel input/output [I/O]).
- Local and distance visualization and data access.
- Comparative visualization and statistical analyses.
- Data set regridding, reprojection, and aggregation.
- New multivariate information visualization techniques that integrate data analytics with interactive displays.
- Support for unstructured grids and non-gridded observational data, including geospatial formats often used for observational data sets.
- Workflow analysis and provenance management.

The test bed analyzes climate model output and verifies it against observed data sets. In general, multi-step coordination of the analysis code will be managed using the automated workflow capabilities of VisTrails, which is a part of UV-CDAT and can control distributed data access and processing. The UV-CDAT analysis code is flexible, adaptable, and scalable, allowing users to add additional analyses and diagnostics at any time. Currently, there are over 1,000 atmospheric and land diagnostics output.

UQ leverages the tools and services being developed for the simulation and UV-CDAT analytical component. They have been extended to support the iterative model runs needed in the UQ analysis process for land only.

Results and Discussion

Scientists need versatile and familiar means of interfacing with their data and model results, and the CSSEF test bed offers a powerful new mechanism, as shown in **Figure 2.** The UI environment incorporates popular thick client toolsets, such as UV-CDAT. It allows end users to search, browse, access, and analyze the archive as if they were on a Web browser. Once data are located, they are downloaded directly to the UV-CDAT local application for further processing. The UV-CDAT package allows users to perform comparative exploratory

analysis and high-end visualization while maintaining the entire provenance and workflow script for data reproducibility.

The prototype test bed is also available via the existing Web interface. This user interface allows users to search and discover scientific data from the entire federated system, browse data collection hierarchies, download data collection files individually or in bulk, sub-set collection files, run model components, track deep storage file download requests, and access user profile information.

While the CSSEF scientific community is studying Web browsers and client analysis tools for broad use, we have noted the limitations of creating repeatable processes and provenance capturing. For such use cases, workflow scripts must be well-defined, repetitive computational tasks that integrate existing applications according to a set of rules. Together, the Web interface UI (working with Globus and AKUNA) and the UV-CDAT GUI will help end users create scripts suitable for submission to a batch queue on a supercomputer or their local system.

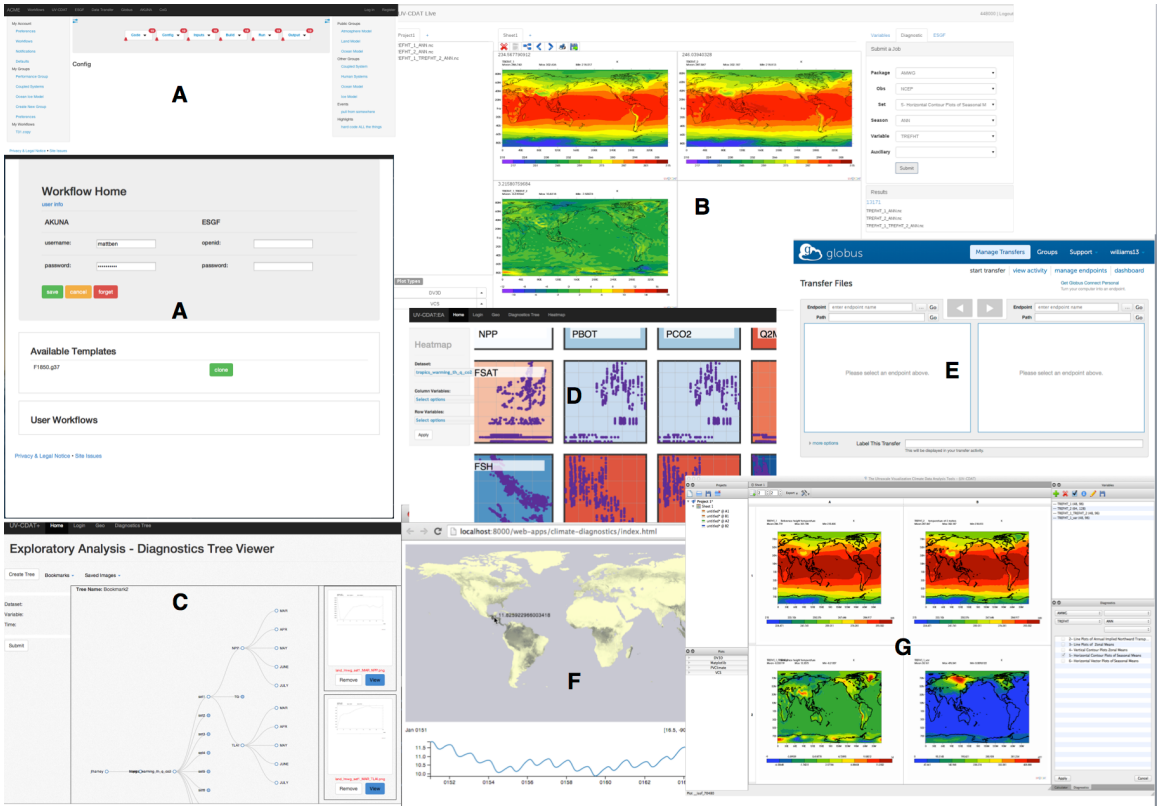

Figure 2. Component images of the CSSEF prototype test bed: (A) The initial prototype test bed UI, callable from any Web browser; (B) UV-CDAT live analysis and diagnostic interactive tool, callable from any Web browser; (c) exploratory diagnostic tree viewer, callable from any Web browser; (D) heat map diagnostics, callable from any Web browser; (E) Globus data transfer tool, callable from any Web browser; and (G) UV-CDAT thick client that is callable from a user's local machine (e.g., desktop, laptop) or high-performance computer

Project Highlights

Test Bed	Prototype test bed infrastructure and software was developed and integrated for all the pieces of the test bed. They include setting up machines, software

	integration between Earth system model code setup, AKUNA running the code on OLCF for the user, message passing from titan to AKUNA to the Web front-end, climatology and diagnostic module set up, automatic data publication to ESGF, UV-CDAT thick and thin client analysis, Web informatics analysis software, and accounts resolution between diverse applications. The team used an agile incremental software development cycle, implemented online full team access tasks and blockers pages, and used Github and git for software repositories.
Messaging	Status messages (using JSON and AMQP) were developed for many of the workflow components to a central tracking authority, while interoperating between Python and Java.
Globus	We integrated Globus GridFTP transfers into the infrastructure to support high-speed, reliable, and secure data movement. We added Globus transfer as an option in the ESGF Web front-end, providing a high-performance managed transfer option for data cataloged and shared using the ESGF software stack. We enhanced the ESGF installer to include steps needed for this capability, and supported deployments in various sites. We began development of the Globus Transfer option in the test bed workflow UI. We integrated Globus identity and group management for user accounts and login. This capability has been integrated with Akuna and the UI, enabling single sign on and capabilities for users to manage multiple accounts.
Model Run	Infrastructure that allows the packaging of the Community Earth System Model (CESM) to be run, authenticate to the compute resources, have the compute resource download, and unpack and take appropriate action. That action can include building and submitting jobs as well as other maintenance tasks.
Cross Machine Job Submission	Data transfers happen automatically for post-compute jobs to speed up the distribution of data and climatologies to be made and diagnostics to be run, as well as other post-processing tasks, across independent machines.
New Diagnostics Framework	New diagnostics framework as an extension of UV-CDAT. For a typical pair of model and observation data sets, over 1000 diagnostics are offered so far. Each diagnostic is a set of plots, typically showing the model and observation data, and their difference, reduced to one or two dimensions. Sometimes, however, the diagnostic data is the result of an involved calculation. Sometimes values of model data are supplemented with their variances. Line and contour plots have been implemented so far. RESTful API wrappers developed for climate diagnostics toolkit. Integrated the uvcmetrics climate diagnostics toolkit project into the framework.
New Diagnostic Plots	New diagnostic 1D, 2D, and 3D plots offered through interactive GUIs, and are computed on the fly. They can be computed either from the raw data or from climatologies (i.e., seasonal time averages). The diagnostics package includes a script for automatically computing climatologies after a model run is completed.
Diagnostics User Interfaces	Two GUIs can run the diagnostics. One is the usual UV-CDAT GUI. With this, one can modify plots interactively with the usual UV-CDAT tools. The other is a Web-based GUI. While UV-CDAT runs on a server, the user needs no software on his remote machine, other than an ordinary Web browser. The Web-based GUI can interactively display three-dimensional variables.

Test Bed User Interface	Designed and implemented the user interface front-end for ACME end-to-end workflows and integration with Globus for the accounts setup, integration with Akuna for model running and workflow provenance, integration with AMQP message passing from titan through Akuna, integration with ESGF publication and Web informatics.
Exploratory Analysis	Exploratory spatio-temporal visualizations; multi-scale correlation heat map; tree-based navigation of diagnostic parameter space; level-of-detail algorithms that allow interactive drill-down of statistical graphics; interactive information visualization techniques using extended parallel coordinates and coordinated views with scatterplots and correlation heatmaps.
EDEN	Visual analytics framework, called the Exploratory Data analysis ENvironment (EDEN), which includes the information visualization techniques and automated data mining algorithms. EDEN is available for download at http://cda.ornl.gov/projects/eden/
AKUNA	Deployed CSSEF AKUNA at ORNL. Implemented a custom job launcher and job handler for the CESM Workflow. Implemented new custom Web services for CSSEF UI to call to launch jobs, retrieve a job's status, list CESM templates, list CESM workflow instances. Implemented API to capture workflow progress and provenance from CESM execution.
ProvEn Services	Integrated AKUNA and ProvEn to facilitate automated provenance capture in CSSEF test bed. System tests are still outstanding.
UV-CDAT	Enhanced the UV-CDAT calculator so it could process normal Python commands as well as variable-targeted operations. The new functionality accurately captures operations on variables, updating the variables panel, as well as allowing basic Python commands in a single interface. Created a common export interface for UV-CDAT spreadsheet cells used for display diagnostic output.

Community Outreach

Report	Third Annual Earth System Grid Federation and Ultrascale Visualization Climate Data Analysis Tools Face-to-Face Meeting Report, March 2014, http://aims-group.github.io/pdf/ESGF_UV-CDAT_Meeting_Report_March2014.pdf
Report	Department of Energy's Biological and Environmental Research Ultra-scale Visualization Climate Data Analysis Tools (UV-CDAT): Three-year Comprehensive Report, November 2013 http://uv-cdat.org/media/pdf/three-year-comprehensive-report.pdf
Paper	"Web-based Visual Analytics for Extreme Scale Climate Science," Supercomputing Conference 2014 (in submission).
Paper	EG Stephan, LK Berg, TO Elsethagen, LD Hogan, MC Macduff, PR Paulson, WJ Shaw, and C Sivaraman, "A Linked Fusion of Things, Services, and Data to Support a Wind Characterization Data Management Facility," *Journal of Web of Things* (in submission).

Paper	Chad A. Steed, Daniel M. Ricciuto, Galen Shipman, Brian Smith, Peter E. Thornton, Dali Wang, and Dean N. Williams. "Big Data Visual Analytics for Exploratory Earth System Simulation Analysis," *Computers & Geosciences*, **61**:71–82, 2013. doi:http://dx.doi.org/10.1016/j.cageo.2013.07.025 CAGEO 2013 Best Paper Award
Paper	Dean N. Williams, Timo Bremer, Charles Doutriaux, John Patchett, Sean Williams, Galen Shipman, Ross Miller, David R. Pugmire, Brian Smith, Chad A. Steed, E. Wes Bethel, Hank Childs, Harinarayan Krishnan, Prabhat, Claudio T. Silva, Emanuele Santos, David Koop, Tommy Ellqvist, Jorge Poco, Berk Geveci, Aashish Chaudhary, Andy Bauer, Alexander Pletzer, Dave Kindig, Gerald L. Potter, and Thomas P. Maxwell. "The Ultra-scale Visualization Climate Data Analysis Tools (UV-CDAT): Data Analysis and Visualization for Geoscience Data," *IEEE Computer*, **46**(9):68–76, 2013. doi:10.1109/MC.2013.119
Paper	Eric G. Stephan, Todd O. Elsethagen, Kerstin Kleese van Dam, and Laura Riihimaki. "What comes first, the OWL or the bean?," Technical Report 790738, figshare, 2013. http://dx.doi.org/10.6084/m9.figshare.790738.
Paper	Stephan, E.G., Pinheiro da Silva, P., and Kleese van Dam, K. 2013. "Bridging the Gap between Scientific Data Producers and Consumers: A Provenance Approach," *Data Intensive Science*. Chapman and Hall/CRC, Boca Raton, FL., 279-299.
Paper	Brian Smith, Daniel M. Ricciuto, Peter E. Thornton, Galen Shipman, Chad Steed, and Dean Williams. "ParCAT: Parallel Climate Analysis Toolkit," in *Proceedings of the International Conference on Computational Science,* pp. 2367–2375, June 2013. doi:10.1016/j.procs.2013.05.408
Paper	Stephan E.G., Elsethagen T.O., Wynne A.S., Sivaraman, C., Macduff, M.C., Berg L.C., and Shaw W.J. "A Linked Fusion of Things, Services, and Data to Support a Collaborative Data Management Facility," Preprints. First International Workshop on Internet Of Things, Austin, TX, 2013.
Paper	Chad A. Steed, Galen Shipman, Peter Thornton, Daniel Ricciuto, David Erickson, and Marcia Branstetter. "Practical Application of Parallel Coordinates for Climate Model Analysis," in *Proceedings of the International Conference on Computational Science*, pp. 877–886, June 2012. doi:10.1016/j.procs.2012.04.094
Poster	Climate Workflow Automation and Optimization on DOE Resources, Science Advisory Meeting, ORNL, March 12, 2014
Poster	Workflow Automation and Optimization on DOE Resources, ORNL Software Expo, May 7, 2014
Poster	Rachana Ananthakrishnan and Ian Foster, "Globus Online: Climate Data Management for Small Teams," AGU Meeting, 2013.