# Visualization and Analysis Tools for Ultrascale Climate Data

PAGES 377–378

Increasingly large climate model simulations are enhancing our understanding of the processes and causes of anthropogenic climate change, thanks to very large public investments in high-performance computing at national and international institutions. Various climate models implement mathematical approximations of nature in different ways, which are often based on differing computational grids. These complex, parallelized coupled system codes combine numerous complex submodels (ocean, atmosphere, land, biosphere, sea ice, land ice, etc.) that represent components of the larger complex climate system.

Climate scientists learn from these simulations by comparing modeled and observed data. A variety of grid schemes and temporal and spatial resolutions makes this task challenging even for small data sets. Recent advances in high-performance computing technologies are enabling the production, storage, and analysis of multiple-petabyte output data sets.

Consolidating interagency efforts, a partnership across government, academic, and private sectors has created a novel system that enables climate researchers to solve current and emerging data analysis and visualization challenges. The Ultrascale Visualization Climate Data Analysis Tools (UV-CDAT) software project (http://uv-cdat.org), started in 2010, uses the Python application programming interface combined with C/C++/Fortran implementations for performance-critical software that offers the best compromise between scalability and ease of use [*Williams et al.,* 2013a]. This software is constantly being updated and improved, with the latest update of UV-CDAT released in October 2014.

## *The UV-CDAT Consortium and Its Goals*

The project team worked closely with current and proposed scientific programs within the U.S. Department of Energy's (DOE) Office of Biological and Environmental Research (BER) and NASA to advance the development of state-of-the-art tools in support of their science missions [*Williams et al.,* 2013b]. The UV-CDAT consortium consists of four DOE laboratories (Lawrence Berkeley National Laboratory, Lawrence Livermore National Laboratory, Los Alamos National Laboratory, and Oak Ridge National Laboratory), NASA, the National Oceanic and Atmospheric Administration (NOAA), two universities (New York University and University of Utah), and two private companies (Kitware and Tech-X).

The consortium's primary goals are to explore and develop software and workflow applications needed to integrate DOE's and NASA's climate modeling and measurements archives; to develop infrastructure for national and international simulation and observation data comparisons; and to deploy a wide range of climate data visualization, diagnostic, model metric, and analysis tools with familiar interfaces for very large, high-resolution climate data sets to meet the growing demands of this data-rich community.

The screen shot of the UV-CDAT application in Figure 1 shows a collage of disparate visualization products, all joined seamlessly under one framework.

## *Need for Parallel Computing and Remote Access to Large Volumes of Data*

As climate models become more complex and output data sets become larger, the steps involving generation, movement, and analysis of model output severely tax serial data processing capabilities. Moreover, support for services that provide remote access to large-scale data is essential as community-wide analysis of model results becomes commonplace and as generalized diagnostics and analysis such as multimodel ensembles become available.

Therefore, from the perspective of large-scale data processing, major efforts from DOE, NASA, and NOAA are being devoted to enabling codes to efficiently output simulation results to parallel disk systems. These efforts include developing analysis tools and workflow patterns to efficiently postprocess large volumes of climate model output for scientific purposes, improving data structures for parallel data processing of batch and interactive processing, and including hardware and networks.

## *UV-CDAT Parallel Computing Tools*

The UV-CDAT system is highly extensible and customizable for high-performance interactive and batch visualization and analysis for climate science and other disciplines of geosciences. For very large (i.e., ultrascale) and complex climate data–intensive computing, UV-CDAT's inclusive framework supports Message Passing Interface (MPI) parallelism as well as task farming and other forms of parallelism.

More specifically, the UV-CDAT framework supports the execution of Python scripts running in parallel using the MPI executable commands and leverages DOE-funded general-purpose, scalable parallel visualization tools such as ParaView and VisIt. This is the first system to be successfully designed in this way and with these features. The climate community leverages these tools and others in support of a parallel client-server paradigm, allowing extreme-scale, server-side computing for maximum possible speedup.

## *Data Storage and Input/Output Needs*

In addition, high-resolution models, ensemble analysis, derived data product generation, and intercomparison of model results and observations require substantial data storage infrastructure in terms of both online data storage (high-performance parallel input/output (I/O) environments) and archival data storage. Managing these data sets—from generation to transformation and fusion to archiving—requires a robust infrastructure

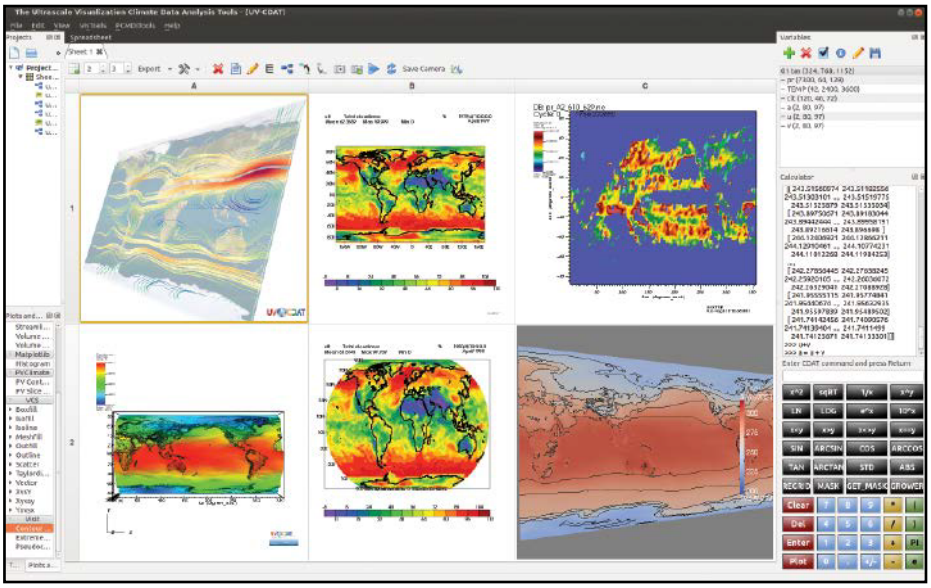

*Fig. 1. Large scale analysis plotting using an array of visualization tools, such as DV3D (top left), VisIt-R plot (bottom left), CDAT plots (middle), VisIt-R (top right), and ParaView (bottom right). Using intuitive drag-and-drop operations, scientists can create, modify, copy, rearrange, and compare visualizations.*

that supports multiple models and storage systems.

State-of-the-art, modern, high-resolution models, such as the DOE Accelerated Climate Modeling for Energy (ACME) project, require parallel I/O environments that provide a structured data model suitable for parallelizing and automating ensemble analysis and intercomparison of models and observational data sets. Observational data sets come in a variety of formats and use numerous metadata models. Supporting the data storage requirements of these data sets—and data products derived from these data sets—requires decoupling the data format from data access and exploratory/parallel analysis.

### *End-to-End Data Management*

UV-CDAT is an integrated framework that provides an end-to-end solution for management, analysis, and visualization of the ultrascale data sets generated for current and future BER climate data repositories and for the climate science community at large. UV-CDAT is based on a client-server architecture and is integrated within the Earth System Grid Federation (ESGF) framework [*Cinquini et al.*, 2013], allowing UV-CDAT to take advantage of the advanced data management mechanisms of ESGF [*Lawrence Livermore National Laboratory,* 2014].

In this way, UV-CDAT provides regridding, reprojection, and aggregation tools directly as a component of the ESGF data node, eliminating or substantially decreasing data movement. The UV-CDAT client provides a turnkey application for building complex data analysis and visualization workflows by interacting with one or more UV-CDAT servers. These workflows may use predefined components for data transformation and analysis, data collection from disparate data sources outside ESGF, visualization, and user-defined processing steps.

### *An Integrated Data Analysis Environment*

The UV-CDAT framework couples powerful software infrastructures through two primary means. One is tightly coupled integration of the CDAT core with the VisTrails/Data Visualization 3D (DV3D)/VTK/ParaView/Earth System Modeling Framework (ESMF) infrastructure to provide high-performance parallel streaming data analysis and visualization of massive climate data sets. The second is loosely coupled integration to provide the flexibility to use tools such as VisIt, Visualization Streams for Ultimate Scalability (ViSUS), R, MATLAB, and ParCat for data analysis and visualization as well as to apply customized data analysis applications within an integrated environment.

To this end, UV-CDAT was designed to incorporate parallel streaming statistics, analysis and visualization pipelines, optimized parallel I/O, remote interactive execution, workflow capabilities, and automatic data provenance processing and capturing. UV-CDAT also offers a novel graphical user interface (GUI; shown in Figure 1) and scripting capabilities for scientists that include workflow data analysis and visualization construction tools as well as the ability to easily add custom functionality. These features are enhanced by the VisTrails provenance tool, the R statistical analysis tool, and advancements in state-of-the-art visualization (DV3D, ParaView, and VisIt), all of which are brought together within a Qt-based GUI.

In the future, UV-CDAT will continue to increase speed, ease of use, and accuracy. In addition, the team is developing interactive capabilities that will enhance diagnostic model capabilities for the ACME project.

### *Acknowledgement*

The developer team consists of Andrew Bauer, Aashish Chaudhary, and Berk Geveci, Kitware, Inc.; Curtis Canada, Phil Jones, and Boonthanome Nouanesengsy, Los Alamos National Laboratory; David Bader, Timo Bremer, Charles Doutriaux, Samuel Fries, Matthew Harris, Elo Leung, Renata McCoy, and Dean N. Williams, Lawrence Livermore National Laboratory; Thomas Maxwell and Gerald Potter, NASA Goddard Space Flight Center; Cecelia DeLuca, Ryan O'Kuinghttons, and Robert Oehmke, National Oceanic and Atmospheric Administration; Ben Burnett, Aritra Dasgupta, Tommy Ellqvist, David Koop, Emanuele Marques, Jorge Poco, Rémi Rampin, Claudio Silva, and Huy Vo, New York University; John Harney, David Pugmire, Galen Shipman, Brian Smith, and Chad Steed, Oak Ridge National Laboratory; and David Kindig and Alexander Pletzer, Tech-X, Inc.

### *References*

Cinquini, L., et al. (2013), The Earth System Grid Federation: An open infrastructure for access to distributed geospatial data, *IEEE Future Gener. Comput. Syst., 36,* 400–417, doi:10.1016/j.future.2013.07.002.

Lawrence Livermore National Laboratory (2014), Third Annual Earth System Grid Federation and Ultrascale Visualization Climate Data Analysis Tools Face-to-Face Meeting report, *Rep. LLNL-TR-650500,* Livermore, Calif. [Available at http://aims-group.github.io/pdf/ESGF_UV-CDAT_Meeting_Report_March2014.pdf.]

Williams, D. N., et al. (2013a), The Ultra-scale Visualization Climate Data Analysis Tools (UV-CDAT): Data analysis and visualization for geoscience data, *Computer, 46*(9), 68–76, doi:10.1109/MC.2013.119.

Williams, D. N., et al. (2013b), Department of Energy's Biological and Environmental Research Ultra-scale Visualization Climate Data Analysis Tools (UV-CDAT): Three-year comprehensive report, U.S. Dep. of Energy, Washington, D. C. [Available at http://uv-cdat.org/media/pdf/three-year-comprehensive-report.pdf.]

—Dean N. Williams, Lawrence Livermore National Laboratory, Livermore, Calif.: email: williams13@llnl.gov