Data Management and Analysis in Support of DOE Climate Science

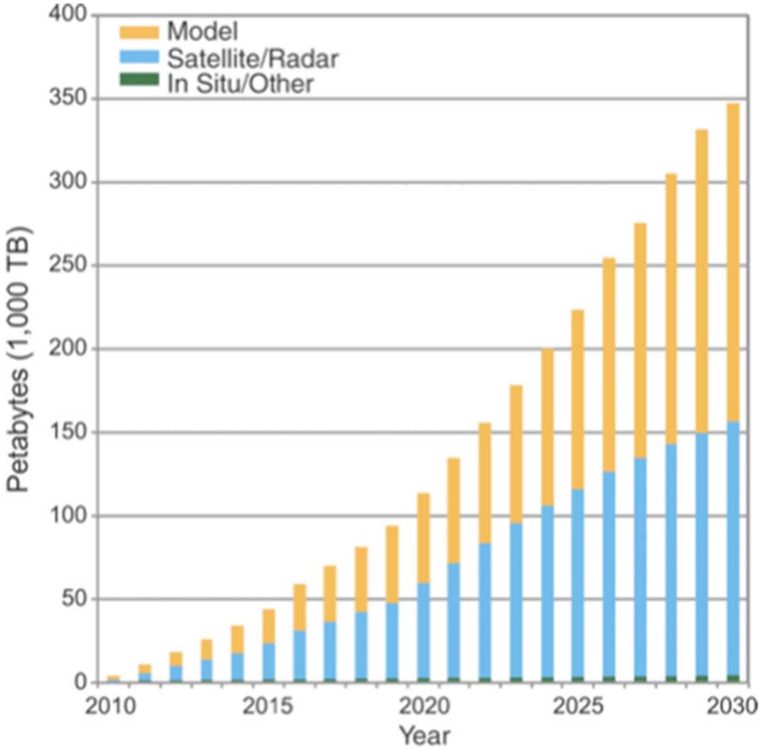
August 7th, 2013

Dean Williams, Galen Shipman

Presented to: Processing and Analysis of Very Large Data Sets Workshop

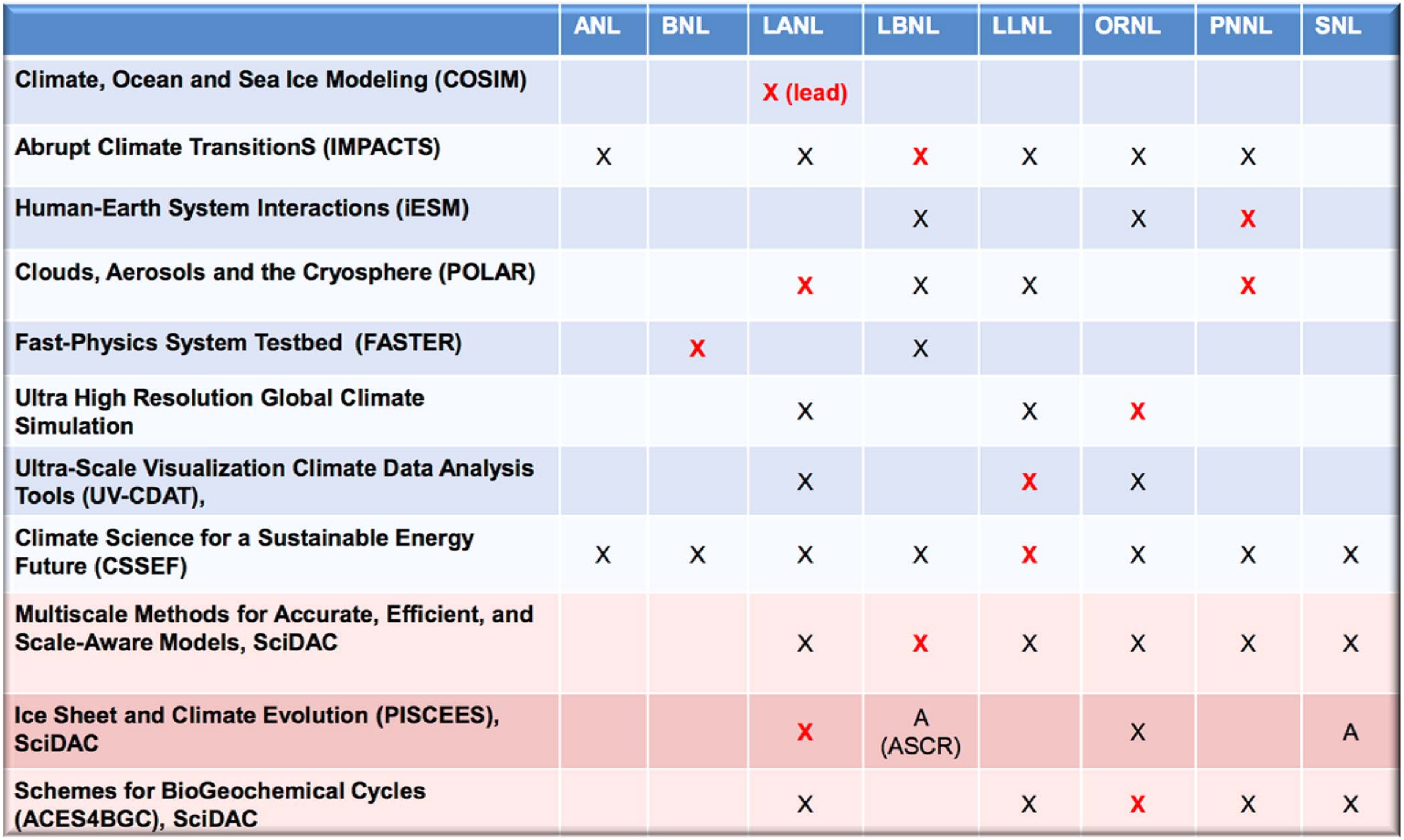
The Climate Data Challenge

- UQ Studies require many model runs for a single study
  - Thousands of single point runs
  - Hundreds of global gridded runs
- Validation of modeled processes
  - Inter comparison with multi-modal observation data
- Increasing spatial and temporal resolution: explore regional scale phenomenon, diurnal cycle
  - Spatial resolution from 1 degree to 0.1 degree
  - Temporal resolution from monthly to hourly



The figure shows the projected increase in global climate data holdings for climate models, remotely sensed data, and in situ.) Science, February 11th 2011

Current DOE Earth System Modeling Lab projects

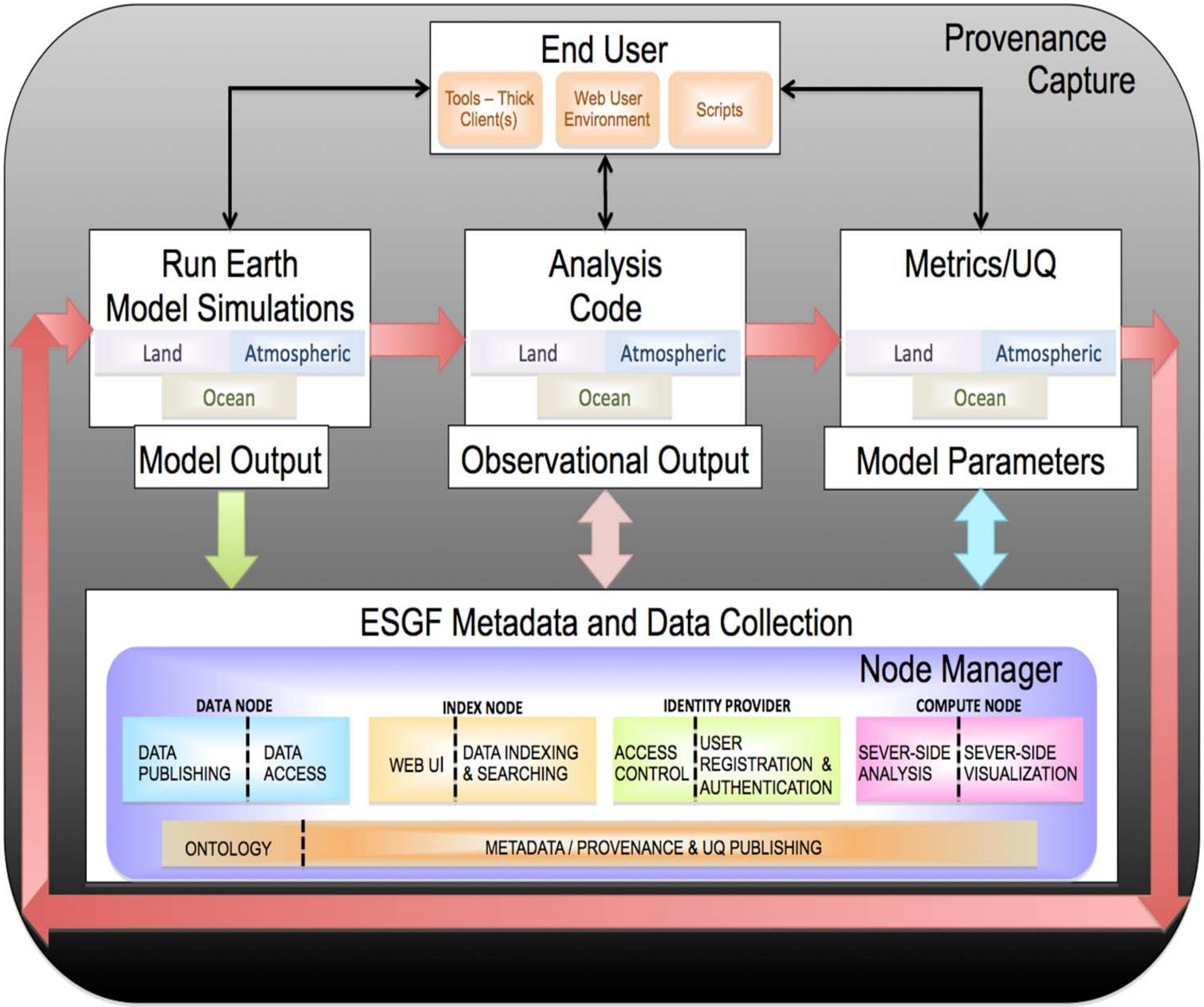


	ANL	BNL	LANL	LBNL	LLNL	ORNL	PNNL	SNL
Climate, Ocean and Sea Ice Modeling (COSIM)			**X (lead)**					
Abrupt Climate TransitionS (IMPACTS)	X		X	**X**	X	X	X	
Human-Earth System Interactions (iESM)				X		X	**X**	
Clouds, Aerosols and the Cryosphere (POLAR)			**X**	X	X		**X**	
Fast-Physics System Testbed (FASTER)		**X**		X				
Ultra High Resolution Global Climate Simulation			X		X	**X**		
Ultra-Scale Visualization Climate Data Analysis Tools (UV-CDAT),			X		**X**	X		
Climate Science for a Sustainable Energy Future (CSSEF)	X	X	X	X	**X**	X	X	X
Multiscale Methods for Accurate, Efficient, and Scale-Aware Models, SciDAC			X	**X**	X	X	X	X
Ice Sheet and Climate Evolution (PISCEES), SciDAC			**X**	A (ASCR)		X		A
Schemes for BioGeochemical Cycles (ACES4BGC), SciDAC			X		X	**X**	X	X

An Exemplar Use Case:
Climate Change Science for a Sustainable Energy Future

- Accelerate the incorporation of new knowledge, including small-scale process data and observations, into climate models.
- Develop new methods for the rapid validation of improved models including quantifying their uncertainty.
- Develop novel approaches to exploit computing at the level of many tens of petaflops in climate models.
- Approach
  - Identify the most important unresolved processes
  - Identify critical underutilized datasets
  - *Develop comprehensive testbeds*
  - Formal incorporation of uncertainty quantification

Climate Model Testbeds – Supporting rapid prototyping, analysis and uncertainty quantification requires an integrated environment



End User

Tools – Thick Client(s)

Web User Environment

Scripts

Provenance Capture

Run Earth Model Simulations

Land

Atmospheric

Ocean

Model Output

Analysis Code

Land

Atmospheric

Ocean

Observational Output

Metrics/UQ

Land

Atmospheric

Ocean

Model Parameters

ESGF Metadata and Data Collection

Node Manager

DATA NODE

DATA PUBLISHING

DATA ACCESS

INDEX NODE

WEB UI

DATA INDEXING & SEARCHING

IDENTITY PROVIDER

ACCESS CONTROL

USER REGISTRATION & AUTHENTICATION

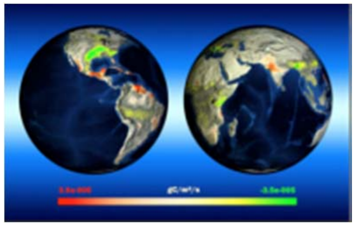
COMPUTE NODE

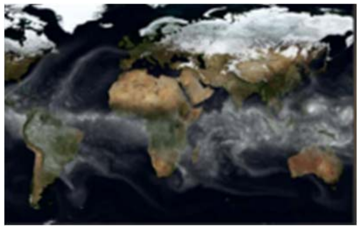
SEVER-SIDE ANALYSIS

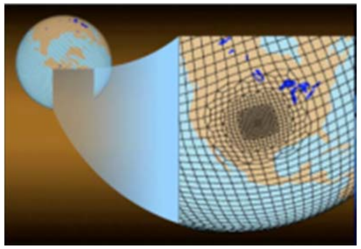
SEVER-SIDE VISUALIZATION

ONTOLOGY

METADATA / PROVENANCE & UQ PUBLISHING

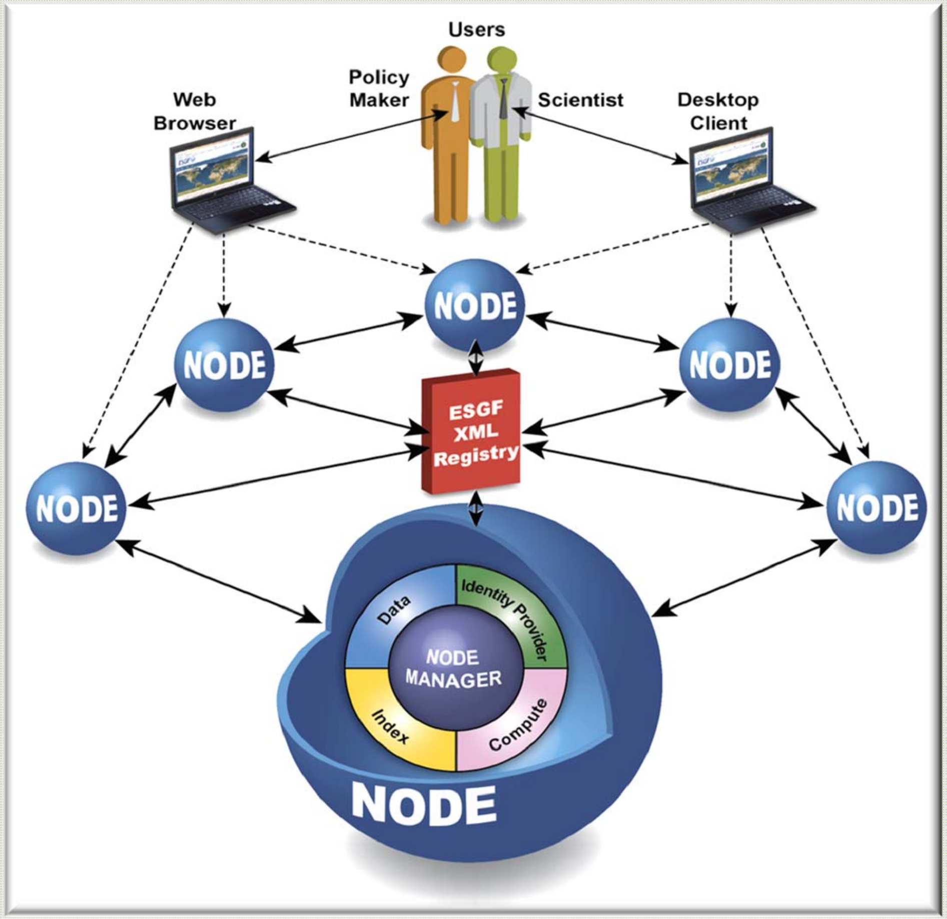






The ESGF distributed data archival and retrieval system

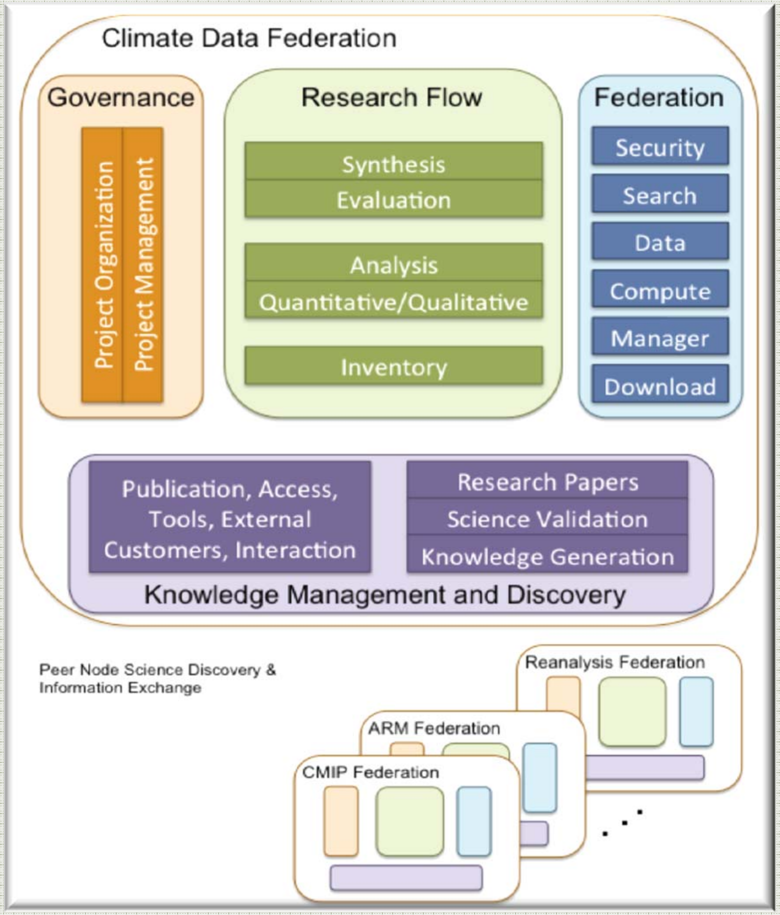
- Distributed and federated architecture
- Support discipline specific portals
- Support browser-based and direct client access
- Single Sign-on
- Automated script and GUI-based publication tools
- Full support for data aggregations
  - A collection of files, usually ordered by simulation time, that can be treated as a single file for purposes of data access, computation, and visualization
- User notification service
  - Users can choose to be notified when a data set has been modified



ESGF software system integrates data federation services

- Publishing
- Search & Discovery
- Archive, Replication, Transport
  - GridFTP, OPeNDAP, DML, Globus Online, ftp, BeSTMan (HPSS)
- NetCDF Climate and Forecast (CF) Metadata Convention
  - (LibCF)
  - Mosaic
- Climate Model Output Rewriter 2 (CMOR-2)
- Regridders: GRIDSPEC, SCRIP, & ESMF
- Data Reference Syntax (DRS)
- Common Information Model (CIM)
- Quality Control
  - QC Level 1, QC Level 2, QC Level 3, Digital Object Identifiers (DOIs)
- Notifications, Monitoring, Metrics
- Security
- Product Services

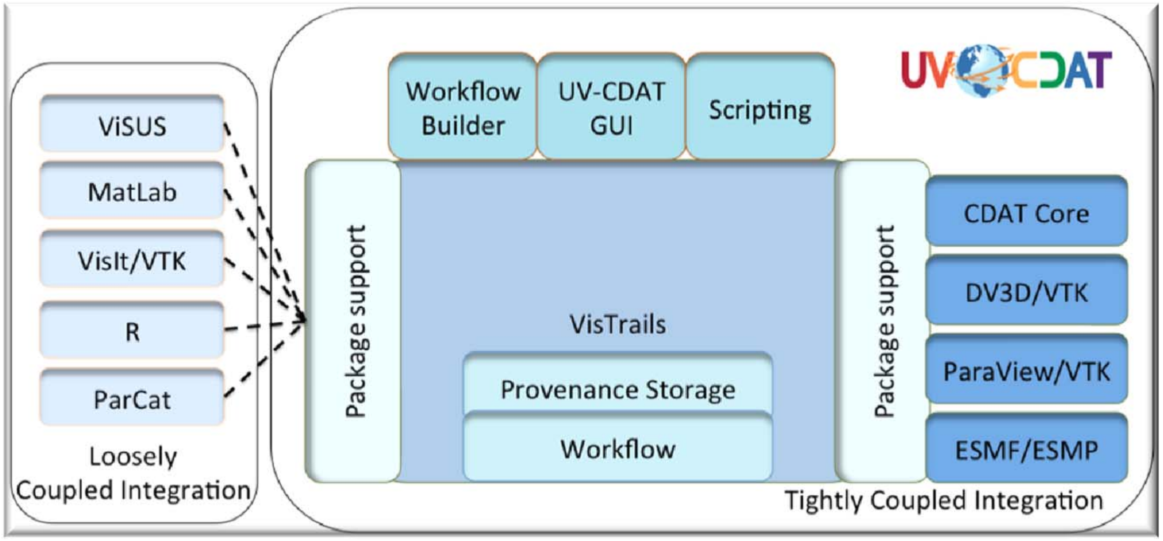
  UV-CDAT, Live Access Server



Ultra-scale Visualization Climate Data Analysis Tools (UV-CDAT)

What is UV--CDAT:

•An integrated environment for data analysis and visualization packages



What is UV-CDAT's purpose:

•Bring together robust tools for climate data processing

•Integration of heterogeneous data sources

•Local and remote data access and visualization

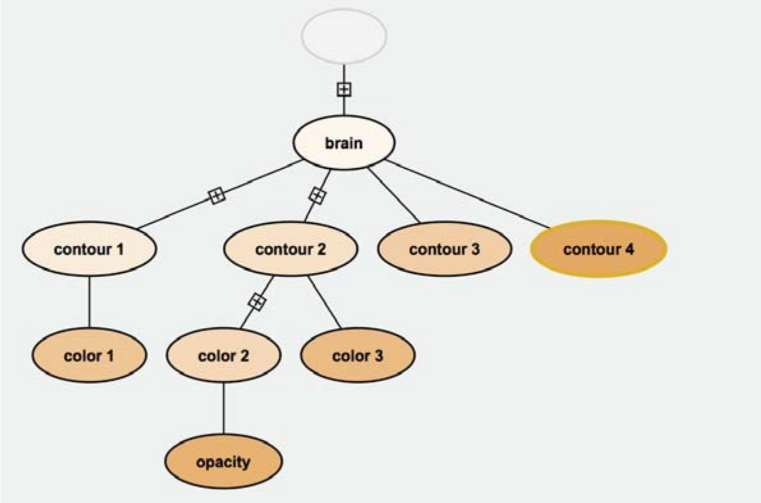
•Reproducibility

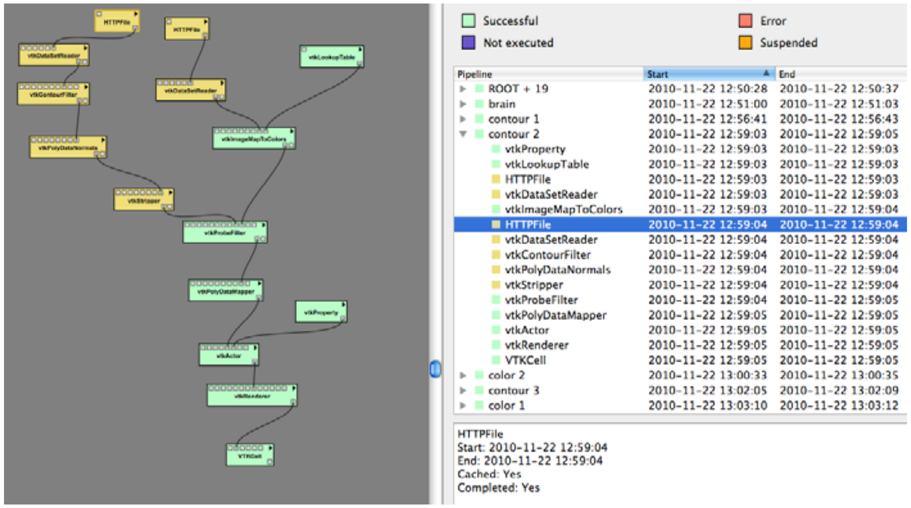
Williams, D.; Doutriaux, C.; Patchett, J.; Williams, S.; Shipman, G.; Miller, R.; Steed, C.; Krishnan, H.; Silva, C.; Chaudhary, A.; Bremer, P.; Pugmire, D.; Bethel, W.; Childs, H.; Prabhat, M.; Geveci, B.; Bauer, A.; Pletzer, A.; Poco, J.; Ellqvist, T.; Santos, E.; Potter, G.; Smith, B.; Maxwell, T.; Kindig, D.; Koop, D., "The Ultra-scale Visualization Climate Data Analysis Tools (UV-CDAT): Data Analysis and Visualization for Geoscience Data," Computer , vol.PP, no.99, pp.1,1, 0
doi: 10.1109/MC.2013.119 URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6487516&isnumber=5306045

UV-CDAT Design Tenant: Workflows and Provenance

- Why use workflows in climate knowledge discovery?
- Integrates multiple tools, languages, and approaches under a unified framework.
- Uses module building blocks to simplify program development.
- Can automate provenance collection.

- Why is provenance important?
- Records all steps in the workflow development and configuration process.
- Records the datasets and parameters used in each KD experiment.
- Records a history of all experiments with results.
- Allows developers to easy back up to an earlier version and start a new branch.



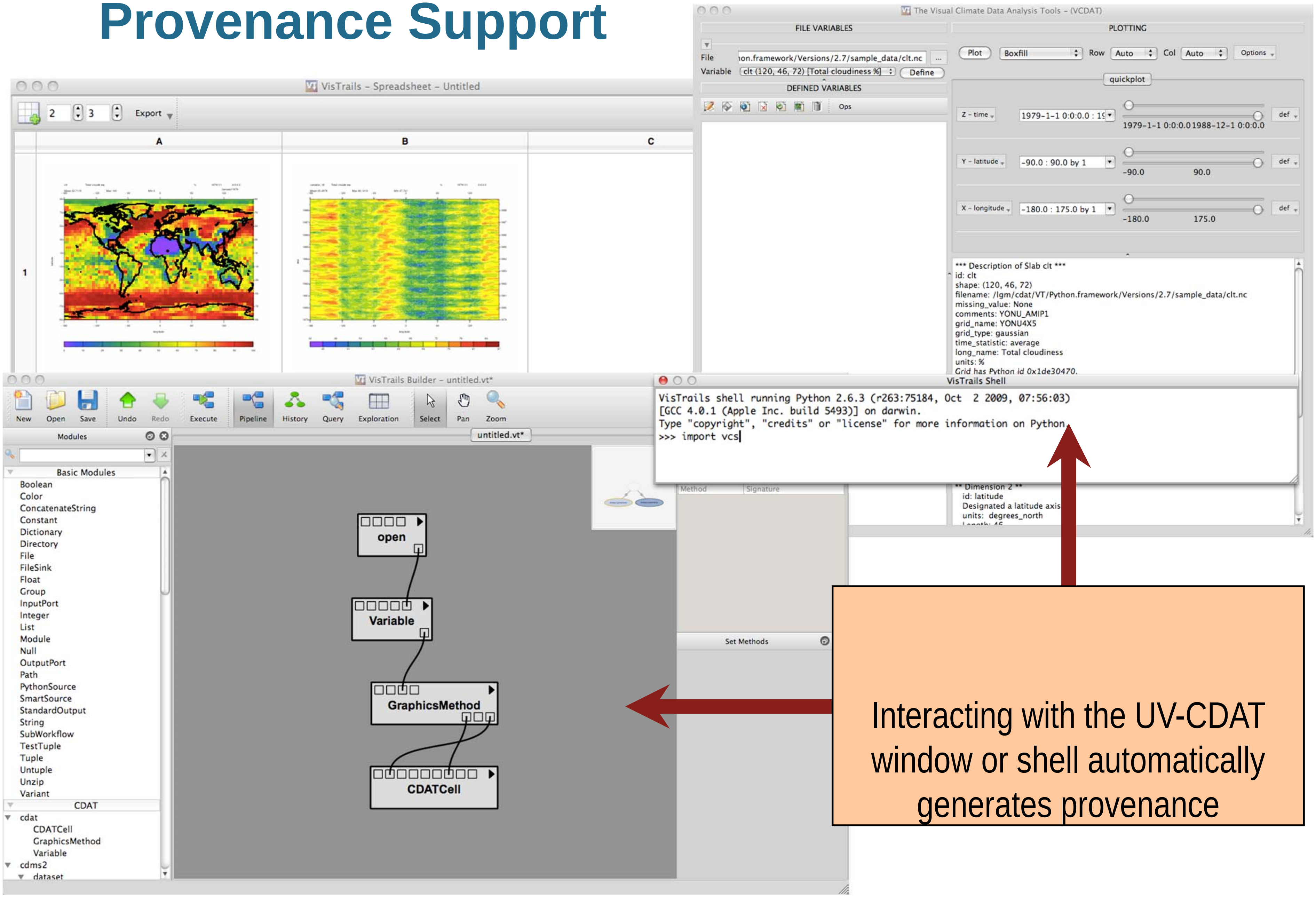


🟩 Successful 🟥 Error
🟪 Not executed 🟧 Suspended

Pipeline	Start ▲	End
▸ 🟩 ROOT + 19	2010-11-22 12:50:28	2010-11-22 12:50:37
▸ 🟩 brain	2010-11-22 12:51:00	2010-11-22 12:51:03
▸ 🟩 contour 1	2010-11-22 12:56:41	2010-11-22 12:56:43
▾ 🟩 contour 2	2010-11-22 12:59:03	2010-11-22 12:59:05
🟩 vtkProperty	2010-11-22 12:59:03	2010-11-22 12:59:03
🟩 vtkLookupTable	2010-11-22 12:59:03	2010-11-22 12:59:03
🟨 HTTPFile	2010-11-22 12:59:03	2010-11-22 12:59:03
🟨 vtkDataSetReader	2010-11-22 12:59:03	2010-11-22 12:59:03
🟩 vtkImageMapToColors	2010-11-22 12:59:03	2010-11-22 12:59:04
🟨 HTTPFile	2010-11-22 12:59:04	2010-11-22 12:59:04
🟨 vtkDataSetReader	2010-11-22 12:59:04	2010-11-22 12:59:04
🟨 vtkContourFilter	2010-11-22 12:59:04	2010-11-22 12:59:04
🟨 vtkPolyDataNormals	2010-11-22 12:59:04	2010-11-22 12:59:04
🟨 vtkStripper	2010-11-22 12:59:04	2010-11-22 12:59:04
🟩 vtkProbeFilter	2010-11-22 12:59:04	2010-11-22 12:59:05
🟩 vtkPolyDataMapper	2010-11-22 12:59:05	2010-11-22 12:59:05
🟩 vtkActor	2010-11-22 12:59:05	2010-11-22 12:59:05
🟩 vtkRenderer	2010-11-22 12:59:05	2010-11-22 12:59:05
🟩 VTKCell	2010-11-22 12:59:05	2010-11-22 12:59:05
▸ 🟩 color 2	2010-11-22 13:00:33	2010-11-22 13:00:35
▸ 🟩 contour 3	2010-11-22 13:02:05	2010-11-22 13:02:09
▸ 🟩 color 1	2010-11-22 13:03:10	2010-11-22 13:03:12

HTTPFile
Start: 2010-11-22 12:59:04
End: 2010-11-22 12:59:04
Cached: Yes
Completed: Yes

Workflow and Provenance Support



File: on.framework/Versions/2.7/sample_data/clt.nc
Variable: clt (120, 46, 72) [Total cloudiness %]

Plot: Boxfill Row: Auto Col: Auto

Z – time: 1979-1-1 0:0:0.0 : 19 — 1979-1-1 0:0:0.0 1988-12-1 0:0:0.0 — def
Y – latitude: -90.0 : 90.0 by 1 — -90.0 90.0 — def
X – longitude: -180.0 : 175.0 by 1 — -180.0 175.0 — def

```
*** Description of Slab clt ***
id: clt
shape: (120, 46, 72)
filename: /lgm/cdat/VT/Python.framework/Versions/2.7/sample_data/clt.nc
missing_value: None
comments: YONU_AMIP1
grid_name: YONU4X5
grid_type: gaussian
time_statistic: average
long_name: Total cloudiness
units: %
Grid has Python id 0x1de30470.
```

```
** Dimension 2 **
  id: latitude
  Designated a latitude axis
  units:  degrees_north
```

```
VisTrails shell running Python 2.6.3 (r263:75184, Oct  2 2009, 07:56:03)
[GCC 4.0.1 (Apple Inc. build 5493)] on darwin.
Type "copyright", "credits" or "license" for more information on Python.
>>> import vcs
```

Interacting with the UV-CDAT window or shell automatically generates provenance

UVCDAT and ParaView integration

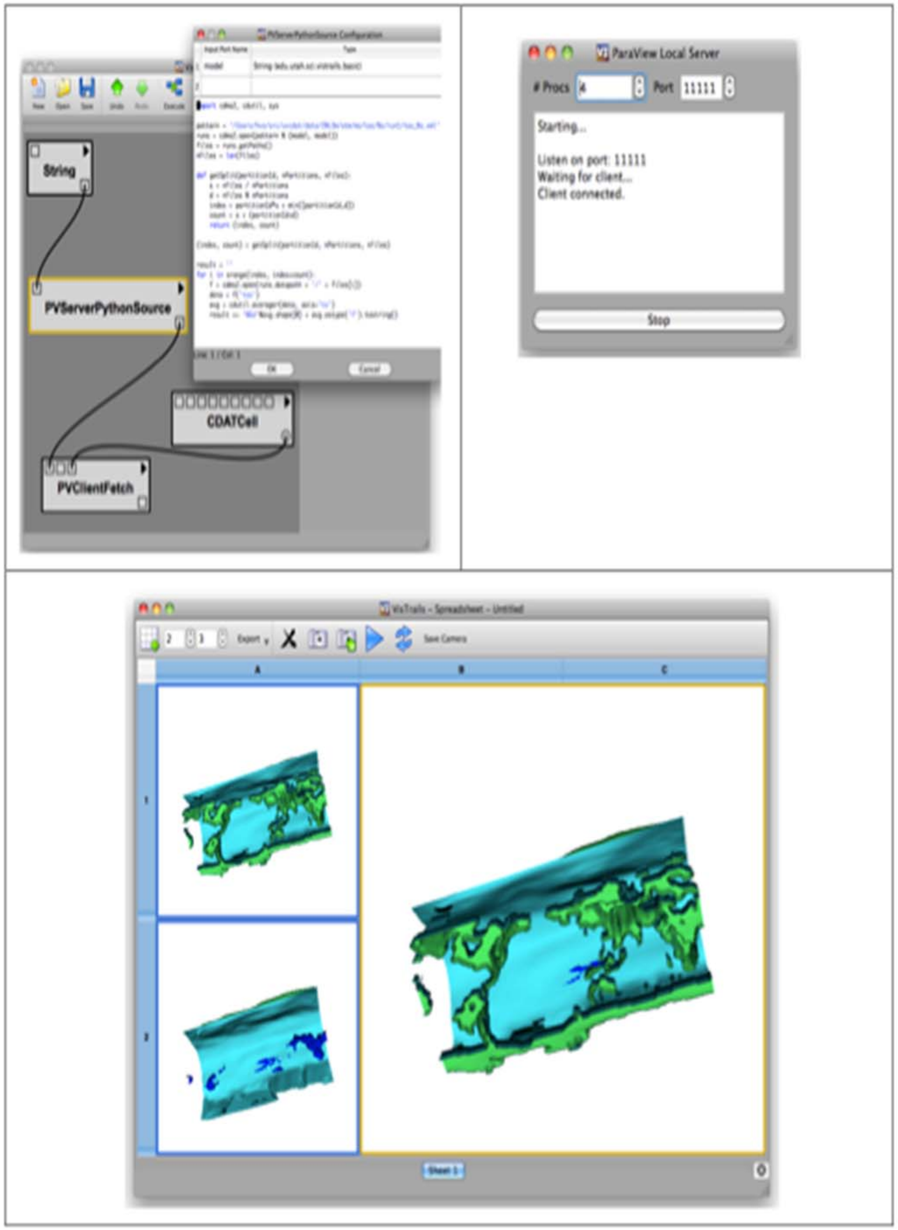
Loosely-coupled workflows executed by ParaViev

Provides Data Parallelism

Remote execution through Paraview Client/Server model

UV-CDAT GUI integration

Ability to access existing CDAT functionality

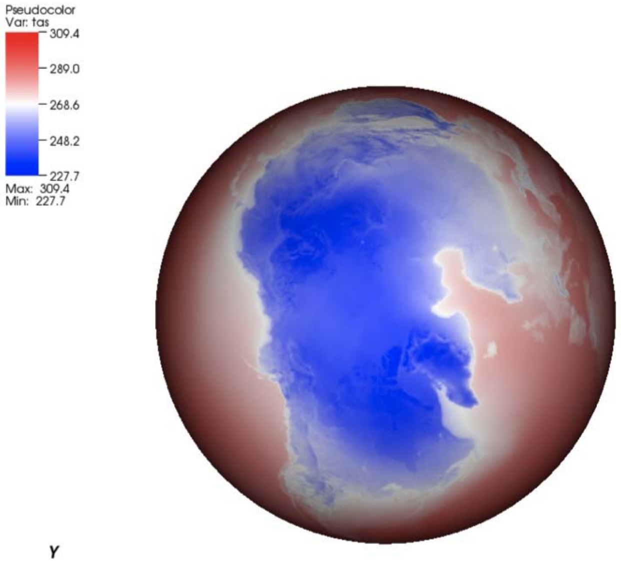


Procs 4 Port 11111

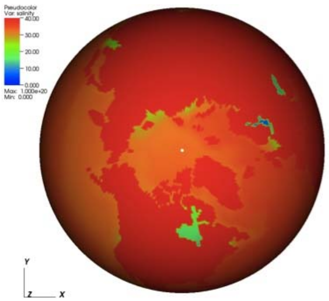
```
Starting...

Listen on port: 11111
Waiting for client...
Client connected.
```

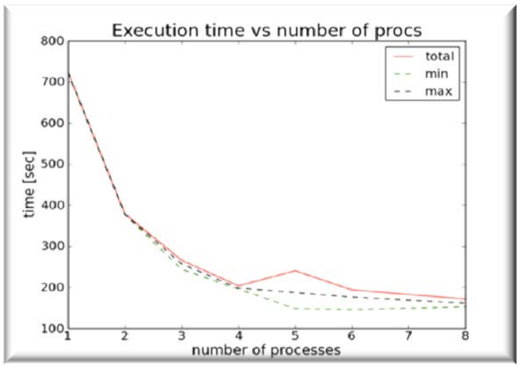
Parallel Processing in UV-CDAT

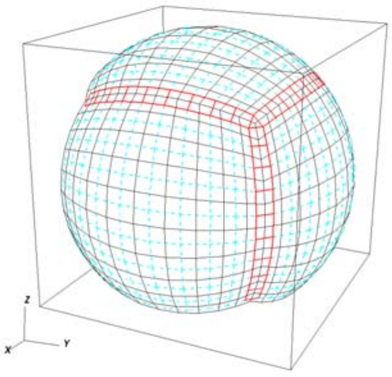
Grid and IO Support



- Parallel I/O Readers
  - Based on NetCDF reader
- Mosaic Grids
  - Supports cubed sphere & tripolar mosaic grids
  - Complies with emerging Gridspec CF standard
- LibCF Interpolation
  - N-Dimensional linear interpolation
  - Curvilinear grids, partial cell masking
- Conservative Interpolation with ESMF
  - Python interface ESMP to Earth System Modeling Framework (ESMF)



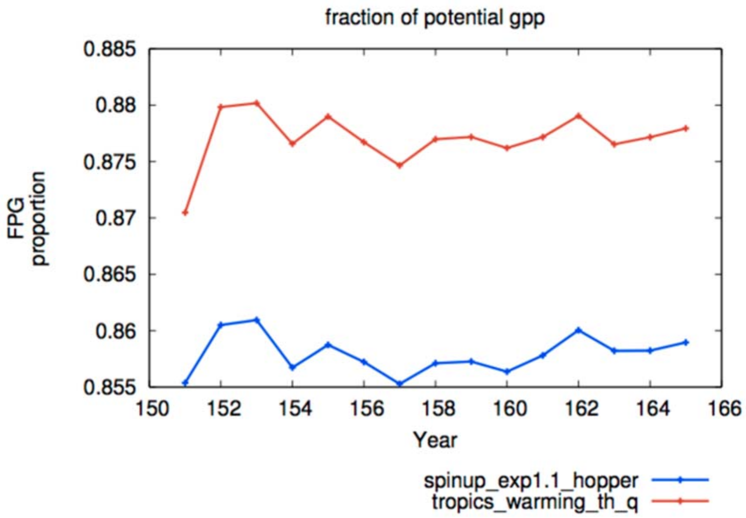




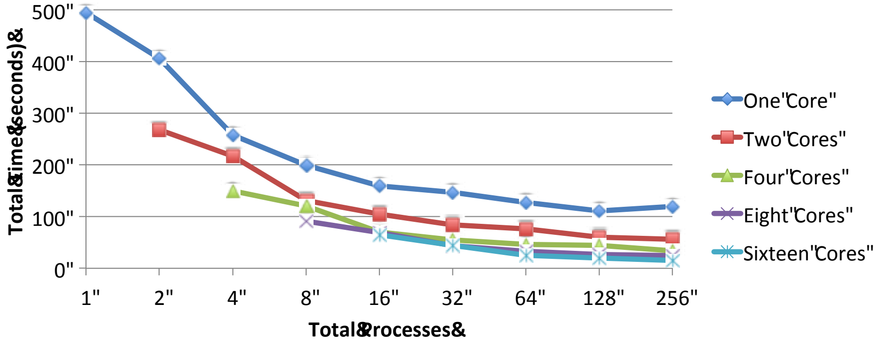
Parallel Processing in UV-CDAT

Highly optimized parallel pre-processing of large scale climate data

- Provides parallel(pre-)processing of large datasets
  - Focus is on datasets with large number of =me steps and/or large number of variables rather than higher spatial resolution
- Calculates point-wise temporal averages (seasonal, monthly, yearly, arbitrary), frequency distributions, and differencing of two datasets (plus difference of averages and average of differences)
- Takes advantage of multicore HPC offerings to improve parallelism and lessen IO issues
- Command-line utility
  - Built upon MPI and parallel NetCDF
- Provides scriptable and embeddable interface
- Buildable and runnable on diverse architectures – laptops to supercomputers
- Reads netCDF input files, produces netCDF outputfiles



Sample plot from ParCAT. Complete analysis package took less than 45 minutes to run compared to over 12 hours using current techniques

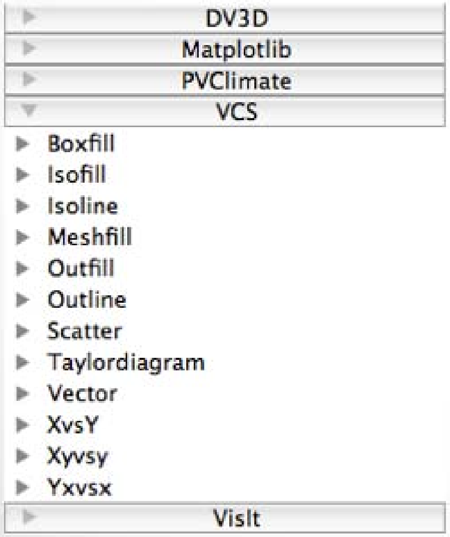


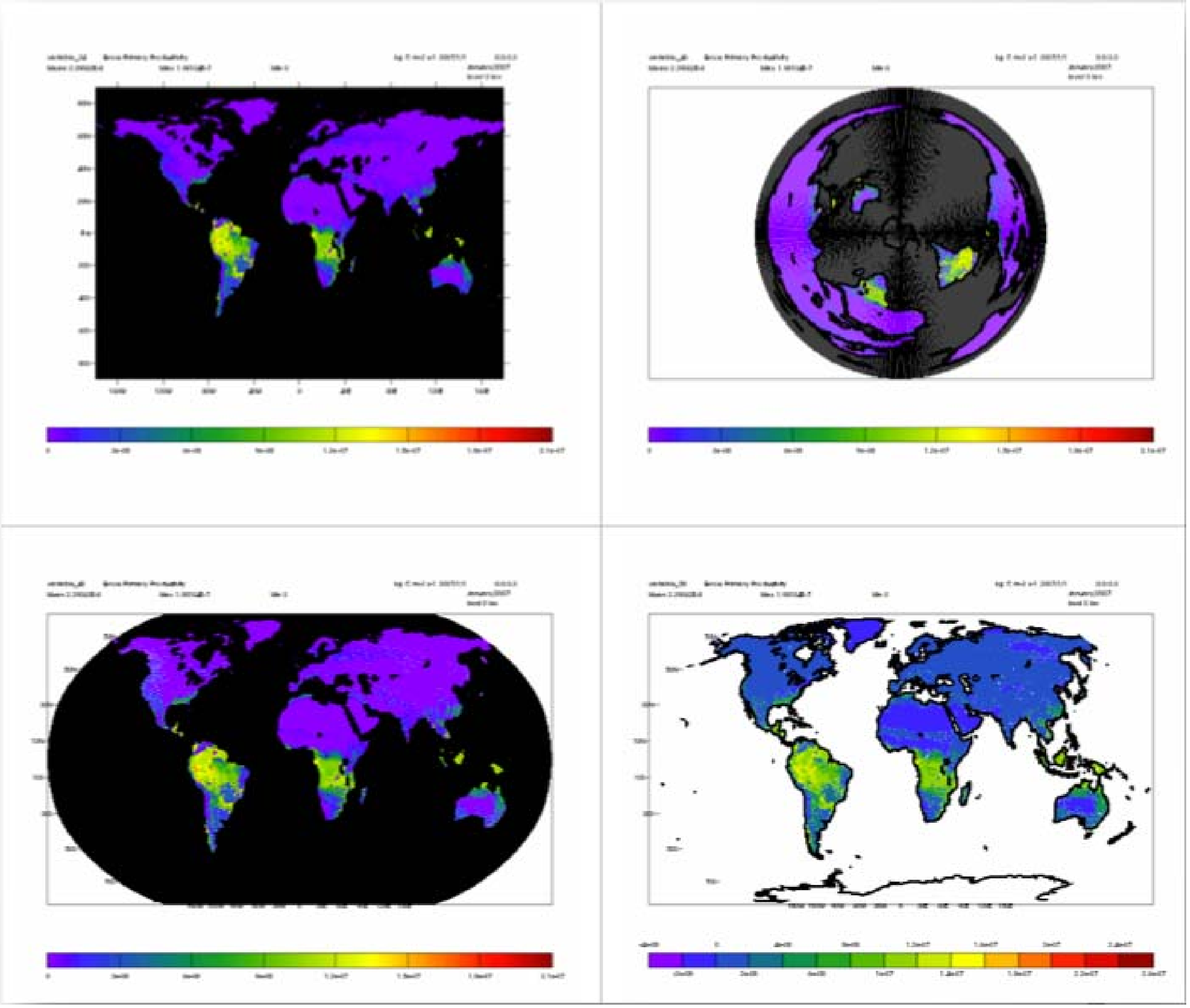
Map Average Time for 349 Variables, CLM Data, 1/2 Degree Resolution, 20 years

Brian Smith, Daniel M. Ricciuto, Peter E. Thornton, Galen Shipman, Chad A. Steed, Dean Williams, Michael Wehner, ParCAT: Parallel Climate Analysis Toolkit, Procedia Computer Science, Volume 18, 2013, Pages 2367-2375, ISSN 1877-0509, http://dx.doi.org/10.1016/j.procs.2013.05.408. (http://www.sciencedirect.com/science/article/pii/S1877050913005516)

CDAT 2D Plots

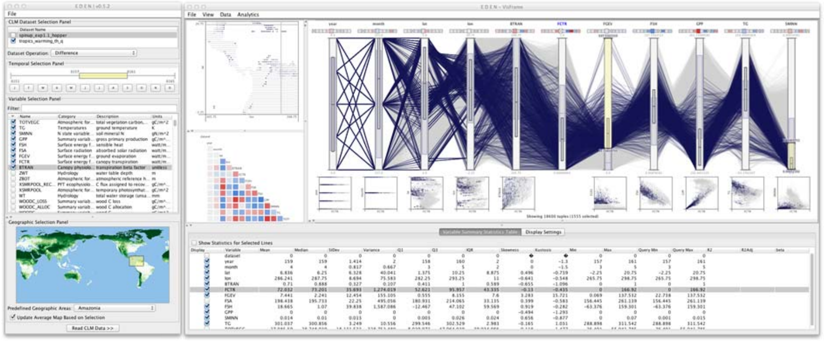
- Many basic plots (boxfill, isofill, isoline, meshfill, etc.)
- Many projections (miller, polar, robinson, etc)



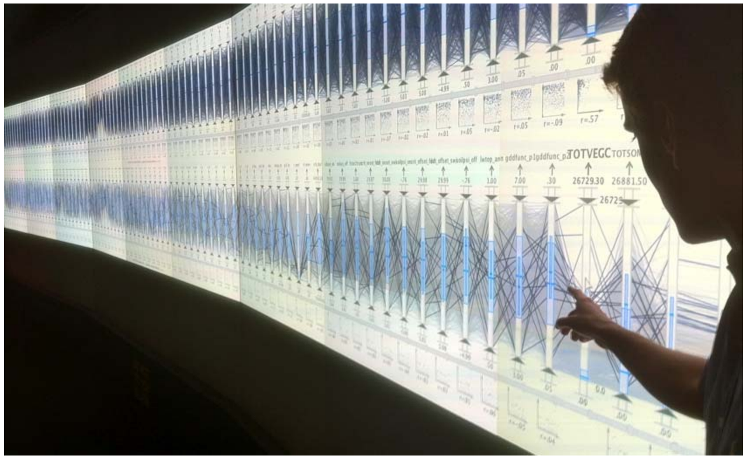


EDEN
Extreme Scale Visual Analytics for Climate Science

- Interactive visual analysis of CLM4 data
  - 100s TBs, 300+ variables, multi-decal, global
- Highlights significant associations to effectively guide the scientist to insight.
- Data summarization for hyper-dimensional data via an intelligent user interface.
- Online linkage to HPC platforms (Titan) for statistical analytics via ParCAT.
- Reduced knowledge discovery timelines.
- Delivered this capability as part of the CLM diagnostics.
- Continued refinement and use by ORNL and NCAR scientists



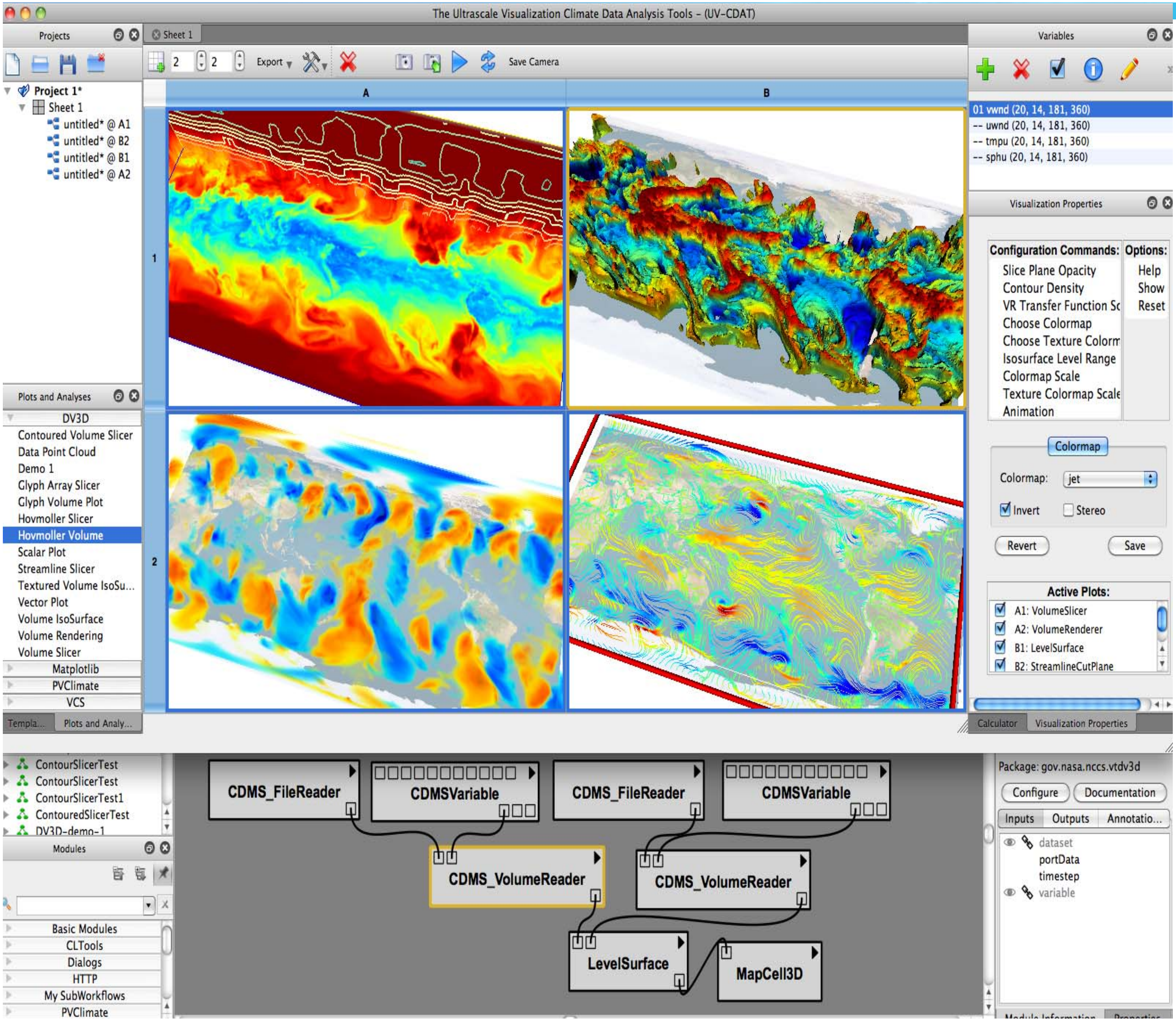
Environmental Data analysis ENvironment (EDEN)



1000 simulations, 81 parameters, 7 output variables
11,520 x 3072 (35 million) pixels

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 Computer Science*, June 2012, pp. 877-886.

DV3D in UVCDAT GUI



01 vwnd (20, 14, 181, 360)
-- uwnd (20, 14, 181, 360)
-- tmpu (20, 14, 181, 360)
-- sphu (20, 14, 181, 360)

Colormap: jet

☑ Invert ☐ Stereo

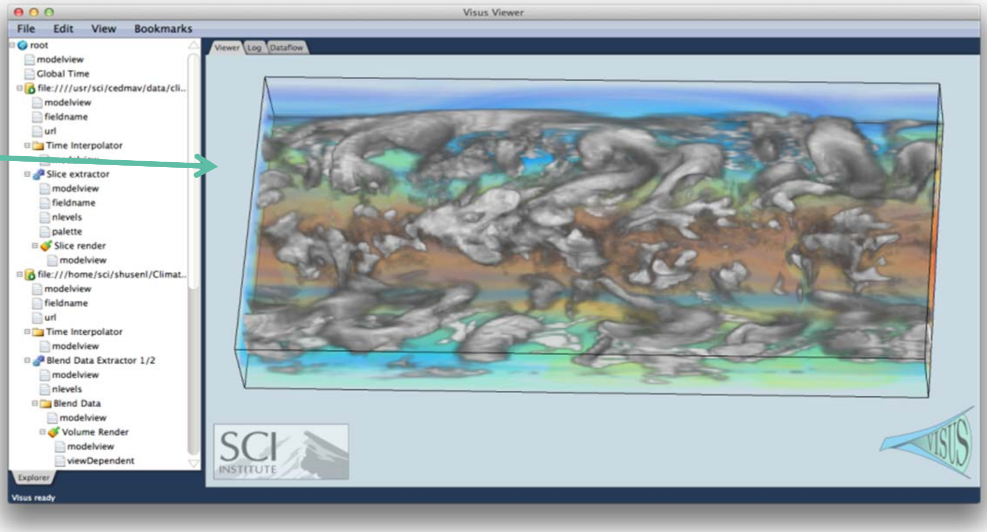
Active Plots:

☑ A1: VolumeSlicer
☑ A2: VolumeRenderer
☑ B1: LevelSurface
☑ B2: StreamlineCutPlane

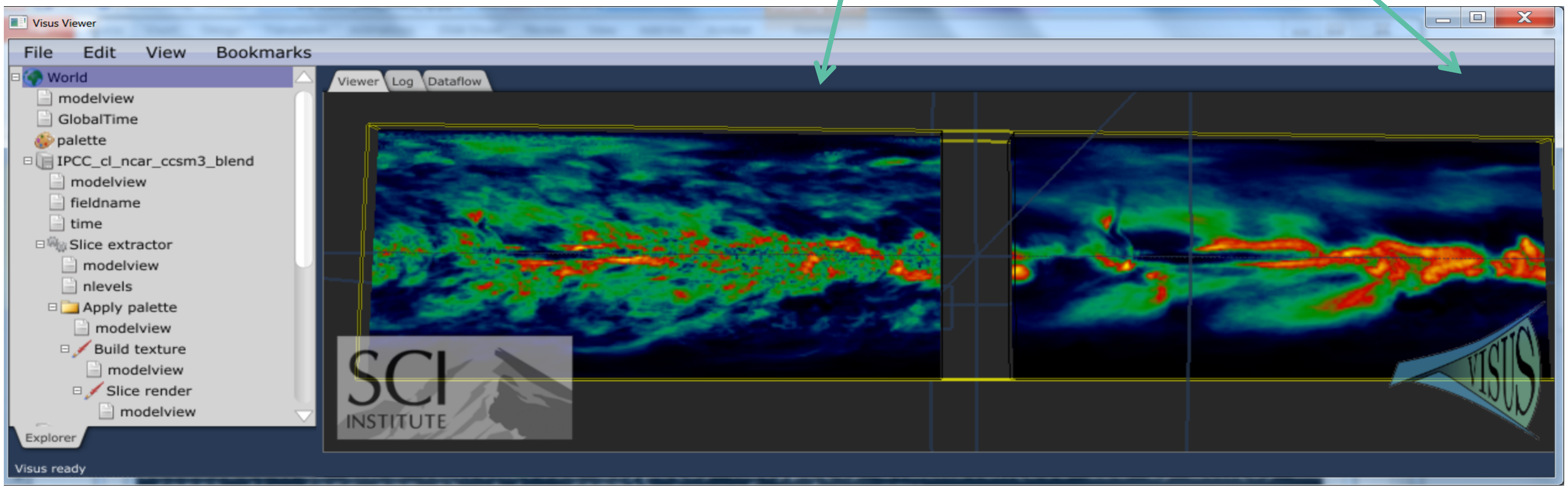
Package: gov.nasa.nccs.vtdv3d

Remote Climate Data Analysis and Visualization

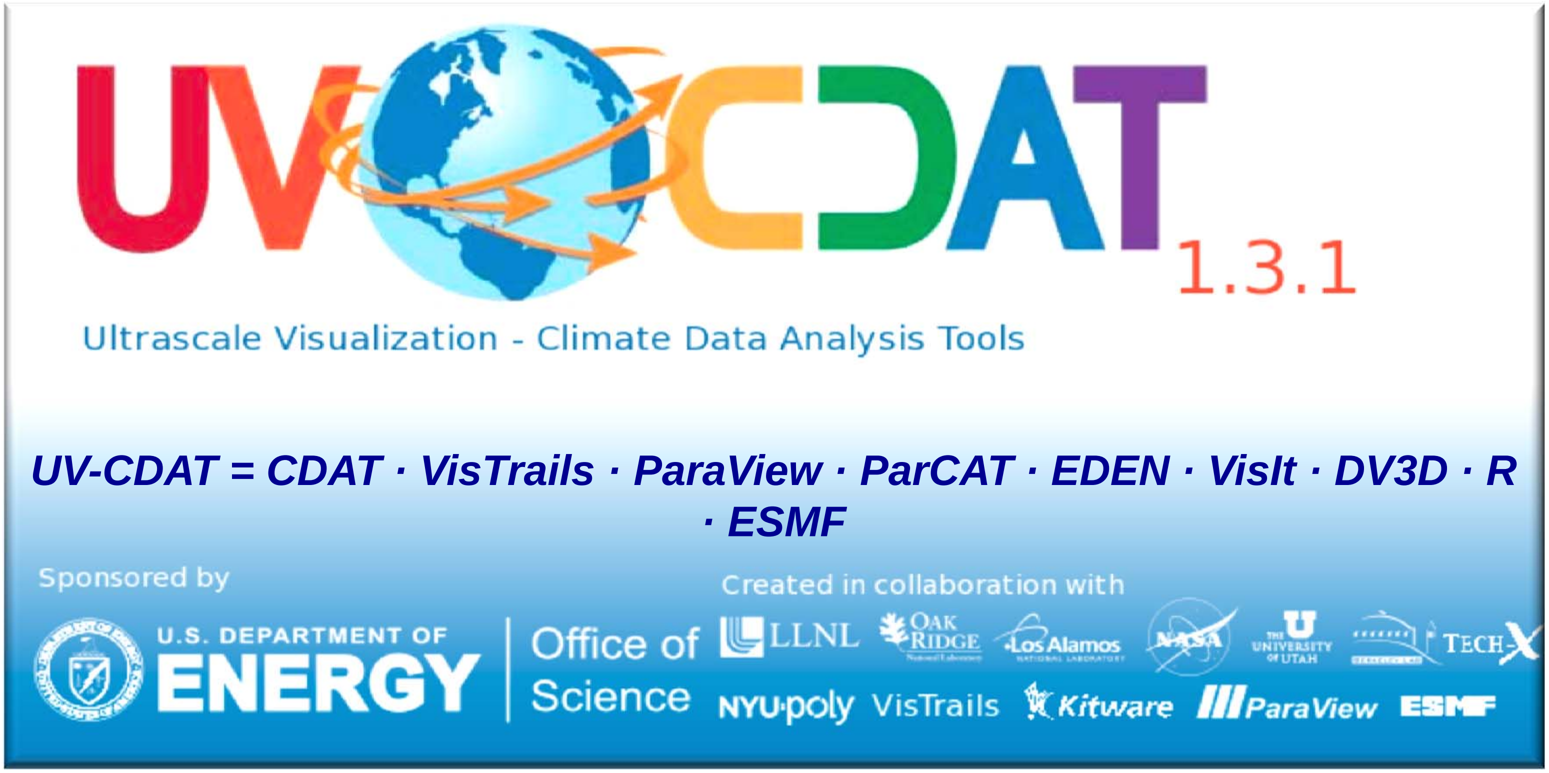
- ViSUS data streams allow merging of multiple datasets in real time
- Time interpolation of and concurrent visualization of climate data ensembles defined on different time scales
- Current development of Qt interface for full integration into UV-CDAT
- Server side and client side computation of statistical functions such as median, average, standard deviation, …….



Standard Deviation and Average of ten climate models



UVCDAT Availability



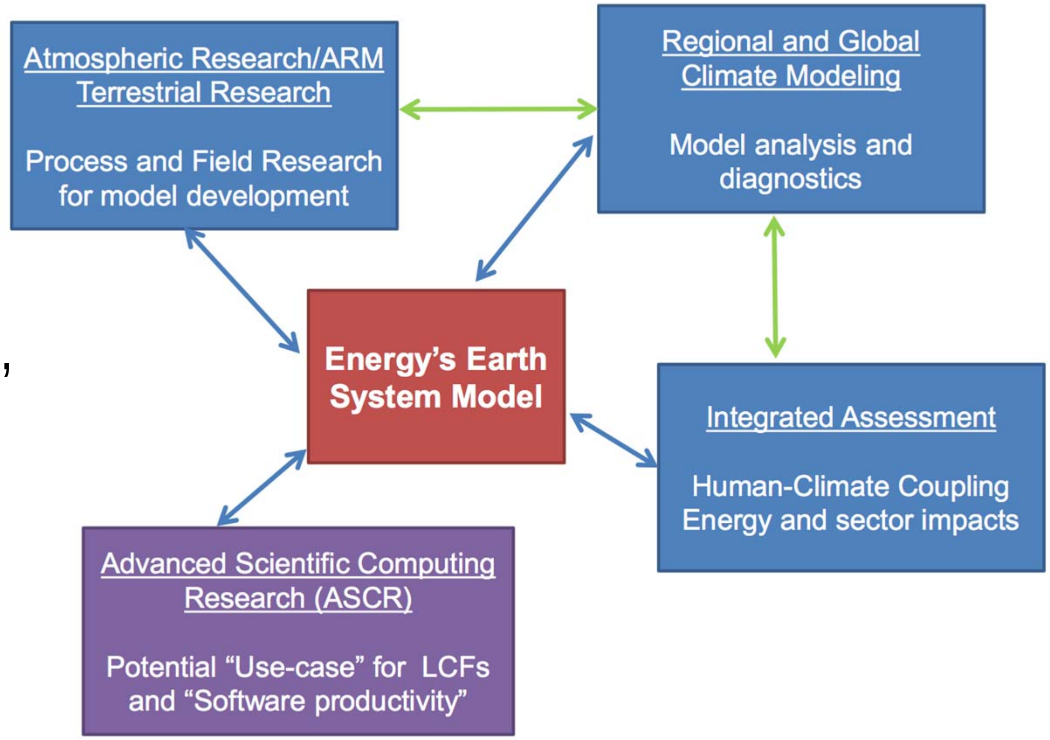
UV-CDAT = CDAT · VisTrails · ParaView · ParCAT · EDEN · VisIt · DV3D · R · ESMF

Release and presentations can be found at the following URLs: *http://uvcdat.llnl.gov/*

User support mailing list: *uvcdat-support@llnl.gov*

Future Direction of Climate Modeling in DOE
- A fully integrated climate modeling program

- Science drivers: hydrological cycle, biogeochemical cycles, and cryospheric systems
- Advance software engineering coding and practice to facilitate automation, calibration, provenance, code performance and code evolution
- Upgrade climate code to efficently utilize current and future DOE Leadership Class Computers
  - Develop climate code architecture that will adapt flexibly to future “extreme-scale computing



Atmospheric Research/ARM
Terrestrial Research

Process and Field Research for model development

Regional and Global Climate Modeling

Model analysis and diagnostics

Energy’s Earth System Model

Integrated Assessment

Human-Climate Coupling
Energy and sector impacts

Advanced Scientific Computing Research (ASCR)

Potential “Use-case” for LCFs and “Software productivity”