

LLNL Triennial Climate Scientific Focus Area Review

Analytics, Informatics, and Management Systems (AIMS): CMIP, ESGF, and other BER Related Data Projects, Software, and Tools

September 5, 2012

Dean N. Williams

On behalf of Multiple Earth System Communities and Projects

Lawrence Livermore National Laboratory



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

LLNL's Analytics, Informatics, and Management Systems (AIMS)" Team (each team member has specific talents and tasks)



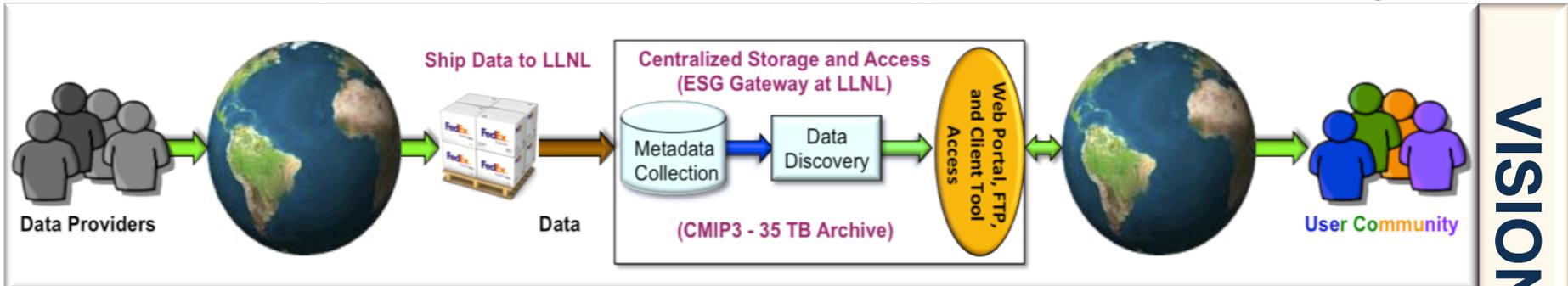
Outline: diverse research interests and coverage

- **Data & Management (Game Changer)**
 - CMIP3 and CMIP5 – Vision for the future
- **Enterprise Software & Systems**
 - **Earth System Grid Federation**
 - Distributed and collaborative workflows, databases, and groupware
 - Large-scale scientific data management
 - Web-based software applications
 - Large-scale analysis and visualization
 - End-to-end simulation and analysis workflow
- **Storage, Movement, and Computational Resources**
 - Computer hardware
 - Computer networks
 - High-performance compute cluster
- **Associated climate and software applications**
- **Current status and near term directions**



Climate research requirements challenge data intensive science and results in a game-changing approach

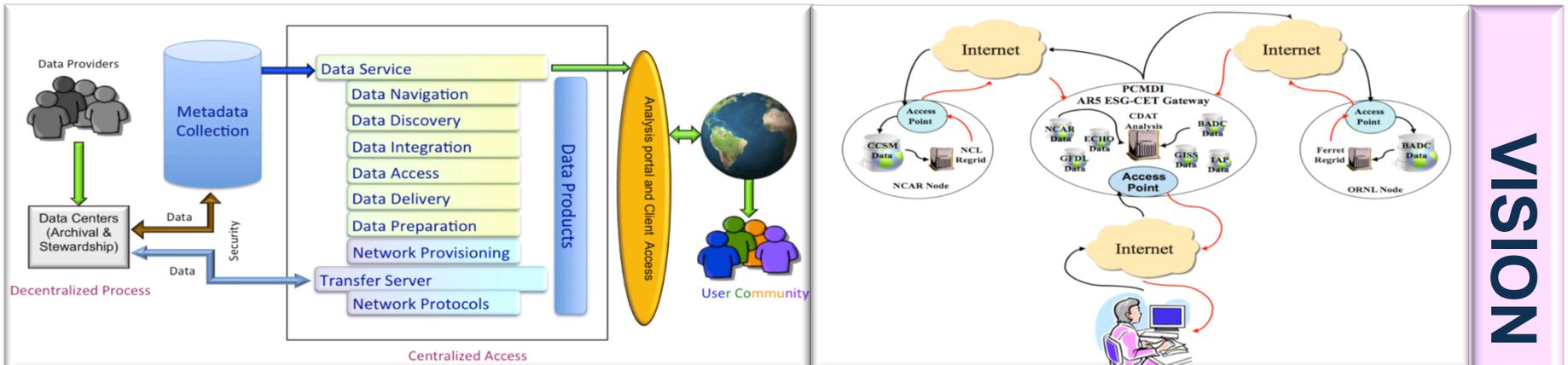
- 2003 centralized processes for CMIP3. CMIP3, data was shipped to LLNL and stored in a single



VISION

“For leadership in implementing, maintaining, and facilitating access to the CMIP3 multi-model data set archive, which led to a new era in climate system analysis and understanding.”—AMS Award 2010

- 2006 decentralized process for CMIP5, a new cloud type approach for distributed data was necessary



VISION

“The new portal is wonderful! In one afternoon work achieved more than several weeks work with the old system.”— Most recent quote from a user 2012

CMIP: experiment design

CMIP5: 47 models **available** from 21 centers

- **CMIP = Coupled Model Intercomparison Project**
 - Phase 1: Idealized simulations of present-day climate (~1 Gigabyte (GB))
 - Phase 2: Idealized simulations of future climate changes (~500 GB: **CMIP2/CMIP1=500**)
 - Phase 3: More realistic simulations (2004 – present) (~35 Terabytes (TB): **CMIP3/CMIP2 = 70**)
- **CMIP 5 multi-model archive expected to include (3.5 Petabytes (PB) **CMIP5/CMIP3 = 100**):**
 - 3 suites of experiments
 - 24 modeling centers in 19 countries
 - 60 models
 - Total data, ~3.5 PB
 - Replica 1 – 2 PB
 - Derived data ~1 PB
- **Global distribution**
- **Timeline fixed by IPCC (2012 - 2013)**
- **LLNL organizes, manages and distributes** the CMIP/IPCC (Intergovernmental Panel on Climate Change) database of climate model output
- **CMIP6 (350 PB – 3 EB ?)**

kilobyte (kB)	10 ³
megabyte (MB)	10 ⁶
gigabyte (GB)	10 ⁹
terabyte (TB)	10 ¹²
petabyte (PB)	10 ¹⁵
exabyte (EB)	10 ¹⁸
zettabyte (ZB)	10 ²¹
yottabyte (YB)	10 ²⁴

Data challenge of CMIP3 archive vs. CMIP5 archive

CMIP3 Modeling Centers		volume (GB)
BCCR	Norway	862
CCCma	Canada	2,071
CNRM	France	999
CSIRO	Australia	2,088
GFDL	USA	3,843
GISS	USA	1,097
IAP	China	2,868
INGV	Italy	1,472
INMCM3	Russia	368
IPSL	France	998
MIROC3	Japan	3,975
MIUB	Germany/Korea	477
MPI	Germany	2,700
MRI	Japan	1,025
CCSM	USA	9,173
UKMO	UK	973
Totals		34,989 (TB)

Archive size: 35 TB

9/5/12

CMIP5 Modeling Centers		volume (TB)
BCC	China	51
CCCma	Canada	51
CMCC	Europe (Italy)	158
CNRM	France	71
CSIRO	Australia	81
EC-EARTH	Europe (Netherlands)	97
GCESS	China	24
INM	Russia	30
IPSL	France	121
LASG	China	100
MIROC	Japan	350
MOHC	UK	195
MPI	Germany	166
MRI	Japan	269
NASA	USA	375
CESM	USA	739
NCC	Norway	32
NCEP	USA	26
NIMR/KMA	Korea	14
NOAA GFDL	USA	158
Totals		3,108 (PB)

**Archive size:
currently: 1.4 PB
total: 3.1 PB by 2013**

CMIP5/CMIP3 = 10²

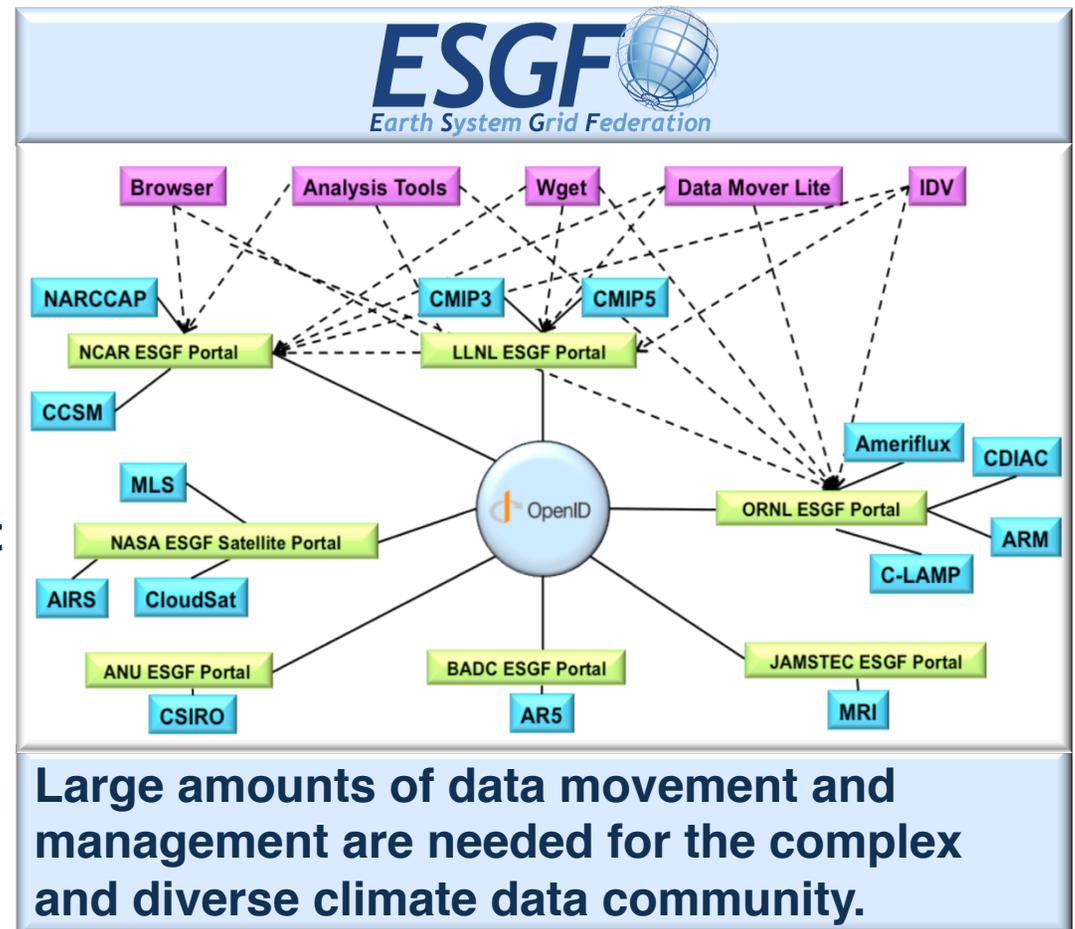
D. N.

A brief history of the Earth System Grid (ESG): ESG-I, ESG-II, ESG-CET, ESGF

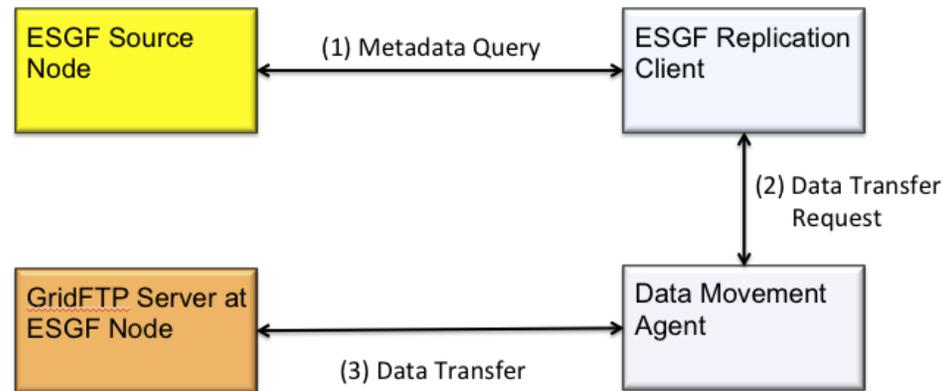
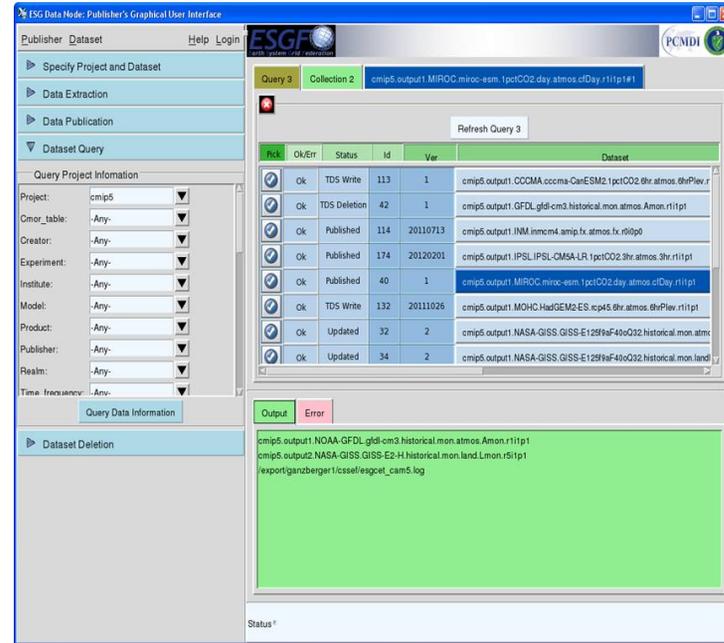
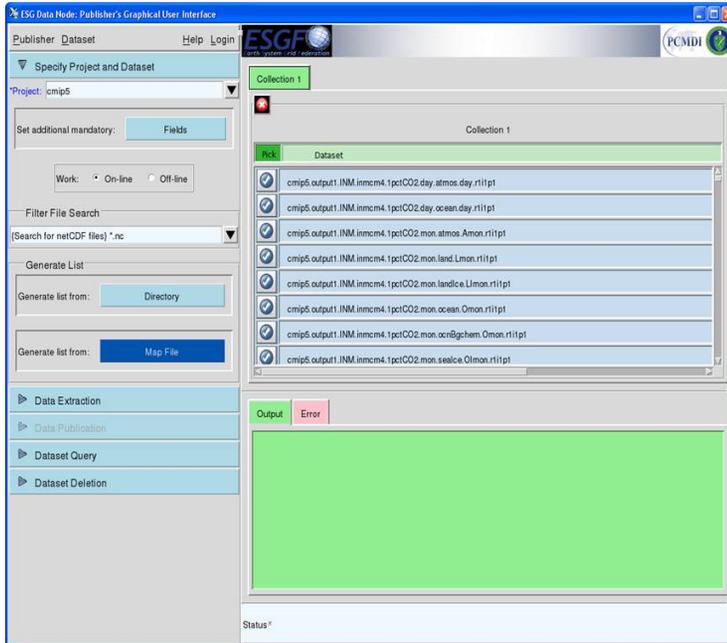
- **ESG-I** funded under DOE' s Next Generation Internet (NGI) to address the emerging challenge of climate data 1999 – 2001 (ANL, LANL, LBNL, LLNL, NCAR, USC/ISI)
 - Data movement and replication; Prototype climate “data browser”; Hottest Infrastructure” Award at SC’ 2000.
- **ESG-II** funded under DOE’ s Scientific Discovery through Advanced Computing (SciDAC), turning climate data sets into community resources 2001-2006 (ORNL addition)
 - Web-based portal, metadata, access to archival storage, security, operational services, 2004 first operational portal CCSM (NCAR), IPCC CMIP3/AR4 (LLNL); 200 TB of data, 4,000 users, 130 TB served.
- **ESG-CET** funded under DOE’ s Offices of ASCR and BER to provide climate researchers worldwide with access to: data, information, models, analysis tools, and computational resources required to make sense of enormous climate simulation and observational data sets 2006 – 2011 (PMEL addition)
 - 2010 Awarded by American Meteorological Society (AMS) for leadership which led to a new era in climate system analysis and understanding.
 - CMIP3, CMIP5, CCSM, POP, NARCCAP, C-LAMP, AIRS, MLS, Cloudsat, etc.
 - 25,000 users, 500-800 users active per month, over 1 PB served
- **ESGF P2P** under the DOE’s Office of BER, it is an open consortium of institutions, laboratories and centers around the world that are dedicated to supporting research of climate change, and its environmental and societal impact. (Additional U.S. funding from NASA, NOAA, NSF.) The federation includes: multiple universities and institution partners in the U.S., Europe, Asia, and Australia.

The ESGF distributed data archival and retrieval system

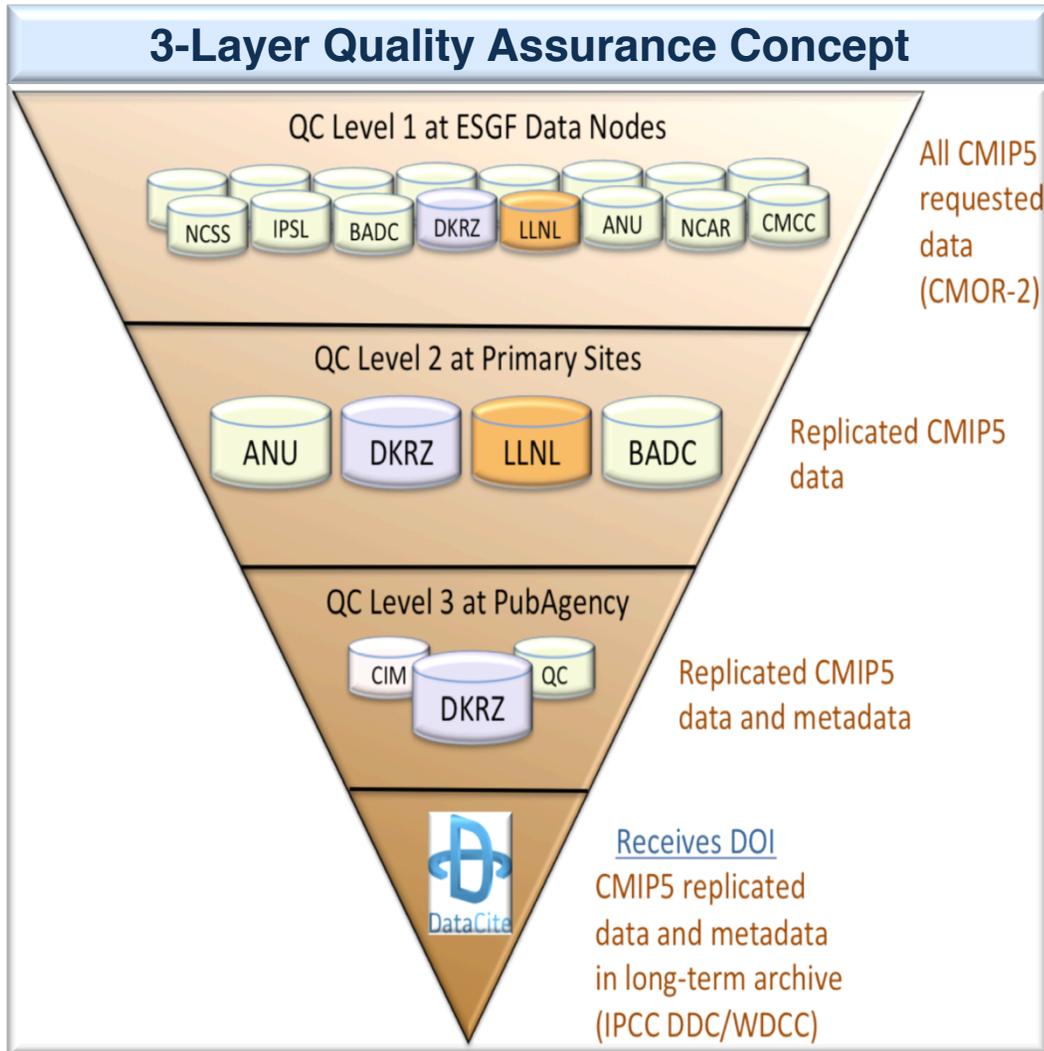
- Distributed and federated architecture
- Support discipline specific portals
- Support browser-based and direct client access
- Single Sign-on
- Automated script and GUI-based publication tools
- Full support for data aggregations
 - A collection of files, usually ordered by simulation time, that can 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



Publication and smart replication



Data quality control check operations end in digital object identifiers (DOIs)



- **Publishing data to an ESGF portal performs QC Level 1 (QCL1) check**
 - QCL1 data are visible to users and are identified as QCL1 on the UI
- **DKRZ (MPI) quality control code is run on data to perform QC Level 2 (QCL2) check**
 - QCL2 data are visible to users and are identified as QCL2 on the UI
- **Visual inspections are performed for inconsistencies and metadata correctness at QC Level 3 (QCL3) check**
 - QCL3 data are visible to users and are identified as QCL3 on the UI
 - **Digital Object Identifiers (DOIs)** are given to data sets that pass the QCL3 check

Publication web application

e.g., <http://cmip.llnl.gov/cmip5/publications>

- Web based submission of citations to database
- Capture tags to enable contextual search, the tags include
 - Keywords
 - Funding
 - Data used (model, experiment, variable, time frequency)
- Submission page differentiate with the journal type
 - The types are: journal, book, proceeding, presentation, technical report
- Editable database entry
 - Enable search of all submitted publications, update, change or add keywords
- Admin page for editing or removing publications

9/5/12

D. N. Williams,

Submit Pages

Search

Review Publications

U.S. DEPARTMENT OF ENERGY Office of Science

CMIP Coupled Model Intercomparison Project
World Climate Research Programme

Submit Edit Publications Administration

Current working build doc

Review Submission

Contact Information: Matthew Harris, 925-423-8978, harris112@llnl.gov
Publication Type: Journal Article
Article Title: The Effects of HTML5 and CSS3 on Front End Load Times
Authors: Matthew B Harris, Renata McCoy
Publication Date: 2012
URL: http://www.mattben.info

Confirm

Send comments. LLNL-WEB-562020, Privacy & Legal Notice

Search Publications

Please enter one or more search criteria below to find the publication records.

Author's Last Name Begins With:

Publication Type(s):

Contact e-mail: *

Year:

Search Reset

CMIP Coupled Model Intercomparison Project
World Climate Research Programme

Submit Edit Publications Administration

Remove	Edit	Author	Article Title	Publication type	Year
<input type="button" value="Remove"/>	<input type="button" value="Edit"/>	Taylor Karl,Stouffer Ronald,Meehl Gerald,	An overview of CMIP5 and the experiment design	J	2012

Go Back and Edit

U.S. DEPARTMENT OF ENERGY Office of Science

CMIP Coupled Model Intercomparison Project
World Climate Research Programme

Submit Edit Publications Administration

All Publications : 2012

Author	Article Title	Journal
Peperov, L. A., Alexandru, R., Laprise, A., Martynov, L., Sushama, ...	Present climate and climate change over North America as simulated by the fifth-generation Canadian Regional Climate Model (CRCM5); (Citation) (More Information)	Climate Dynamics
Ahlström A., G. Schurgers, B. Smith	Robustness and uncertainty in terrestrial ecosystem carbon response to CMIP5 climate change projections; (Citation) (More Information)	Environmental Research Letters
Anav A., P. Friedlingstein, M. Kidston, L. Bopp, P. Clais, ...	EVALUATING THE LAND AND OCEAN COMPONENTS OF THE GLOBAL CARBON CYCLE IN THE CMIP5 EARTH SYSTEM MODELS; (Citation) (More Information)	Journal of Climate
Andrews, J. M., Gregory, M. J., Vellinga, K. E. T.	Forcing, feedbacks and climate sensitivity in Coupled atmosphere-ocean climate models; (Citation) (More Information)	Geophysical Research Letters
Andrews, J. M., Gregory, M. J., Vellinga, K. E. T., ...	Forcing, feedbacks and climate sensitivity in Coupled atmosphere-ocean climate models; (Citation) (More Information)	Geophysical Research Letters

(More Information)

Experiments	Models	Variables	Keywords
abrupt4xCO2	CanESM2	land area fraction	WG1 (physical climate system)
piControl	CNRM-CM5	surface temperature	Abrupt change
sstClim	CSIRO-	toa_incoming_shortwave_flux	

Publication Years

2007	1
2008	2
2011	5
2012	208

CMIP5 QC status

- **QC Status of CMIP5 as of 08/2012:**

- **Quality Control 1:** 54,269 data sets
- **Quality control 2:** 10,986 data sets (finalized 6535)
- **Quality Control 3:** 1468 data sets
- **DataCite DOI:** 1468 data sets

For more information on QC Status, see the following URLs:

<https://redmine.dkrz.de/collaboration/projects/cmip5-qc/wiki>

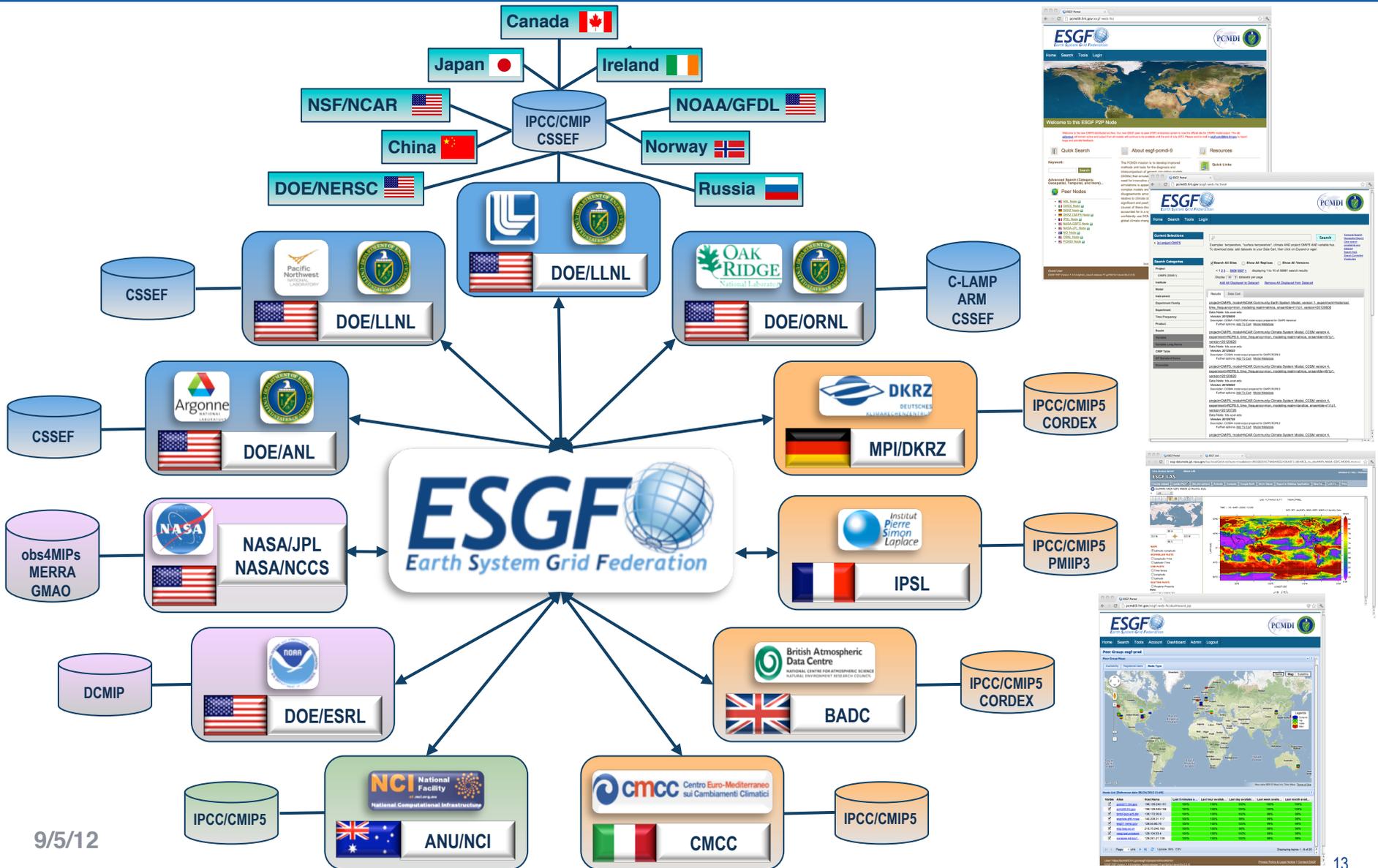
<http://cera-www.dkrz.de/WDCC/CMIP5/QCResult.jsp>

- **Replicated CMIP5**

- **LLNL replicated data sets** 23,745 data sets
- **DKRZ replicated data sets** 23,745 data sets

ESGF is more than CMIP: federated and integrated data from multiple sources

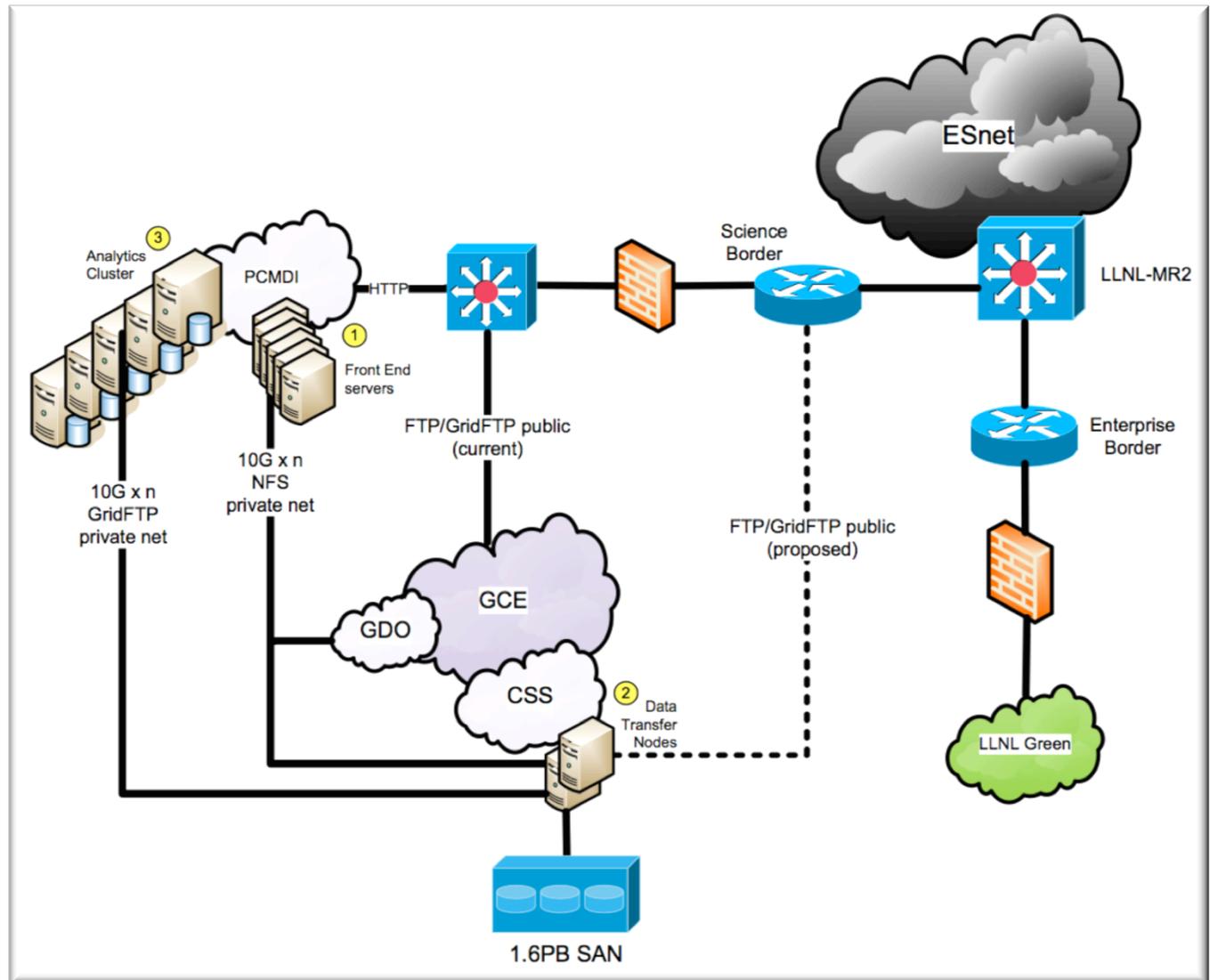
<http://pcmdi9.llnl.gov/esgf-web-fe/>



9/5/12

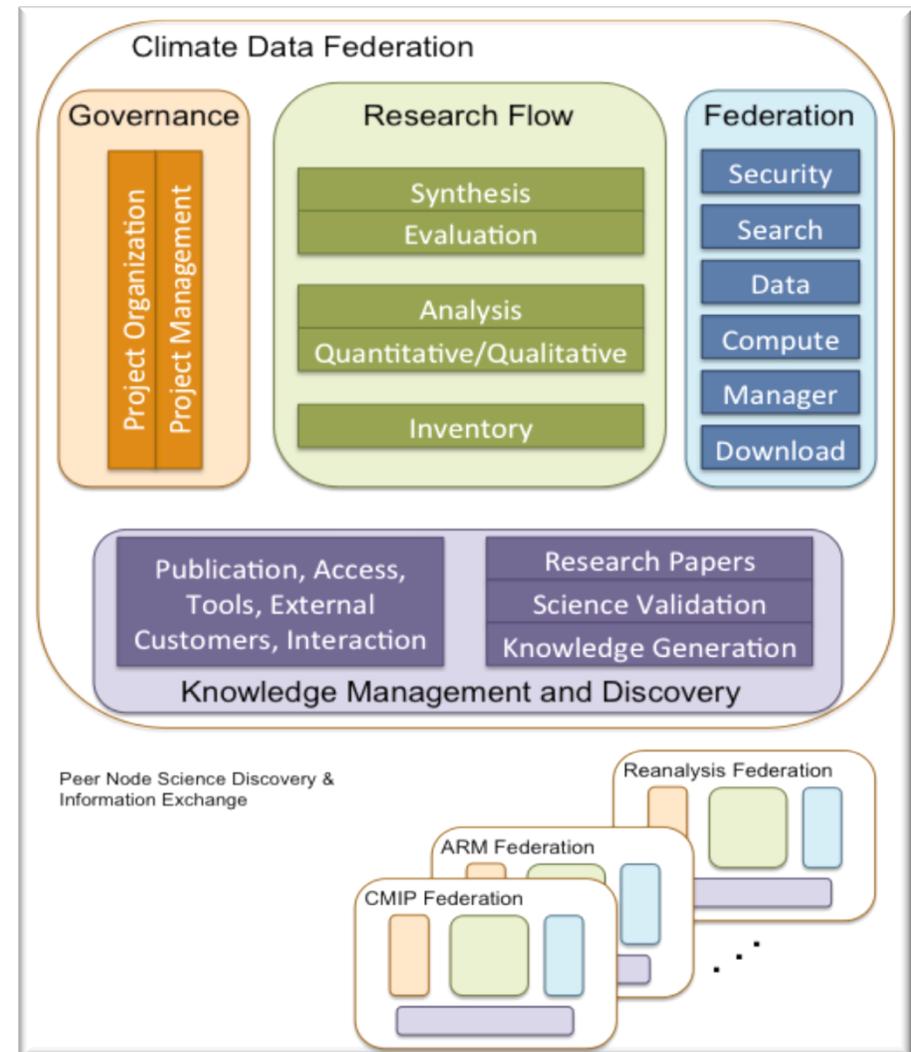
ESGF integrated with the hardware and network

- 1 Users communicate with ESGF front-end servers via HTTP
- 2 Large data sets are made available to users directly from the Climate Storage System (CSS) via vsftp and GridFTP
- 3 Through UV-CDAT, ESGF will perform analysis of raw data if requested by users through the front-end servers to the analysis (hadoop) cluster



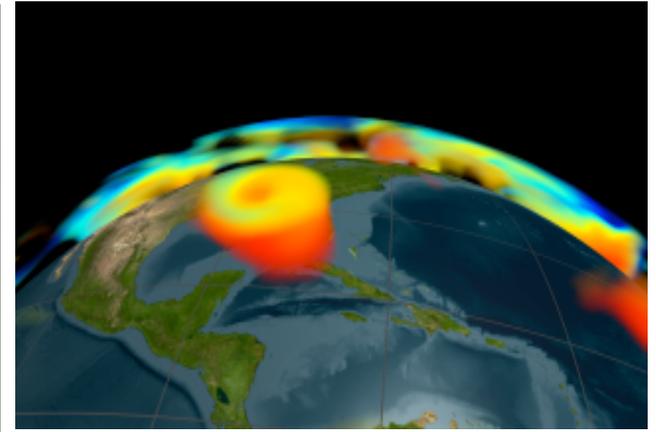
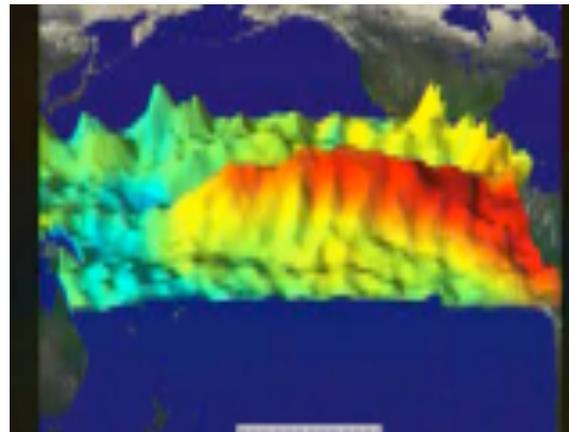
ESGF software system integrates data federation services

- NetCDF Climate and Forecast (CF) Metadata Convention
 - (LibCF)
 - Mosaic
- Climate Model Output Rewriter 2 (CMOR-2)
- Regriders: GRIDSPEC, SCRIP, & ESMF
- Publishing
- Search & Discovery
- Replication and Transport
 - GridFTP, OPeNDAP, DML, Globus Online, ftp, BeSTMan (HPSS)
 - Networks
- Data Reference Syntax (DRS)
- Common Information Model (CIM)
- Quality Control
 - QC Level 1, QC Level 2, QC Level 3, Digital Object Identifiers (DOIs)
- Websites and Web Portal Development
 - Data, Metadata, Journal Publication Application
- Notifications, Monitoring, Metrics
- Security
- Product Services
 - Live Access Server, UV-CDAT



Advanced analytics, informatics, and visualization for scientists

- Analysis and visualization is a key aspect of scientific analysis and discovery
- Advanced interactive visualization is rarely used by scientists
- Interfaces too complex, pickup too costly
- Interactive climate visualization requires:
 - Intuitive interfaces
 - Seamless integration with high performance analysis workbenches
 - Parallel streaming visualization pipelines



Supported data activities (complementary BER funded project)

<http://uvcdat.llnl.gov>

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

- Integrate DOE's climate modeling and measurements archives
- Develop infrastructure for national and international model/data comparisons
- Deploy a wide-range of climate data visualization, diagnostic, and analysis tools with familiar interfaces for very large, high resolution climate data sets (CDAT, VTK, R, VisIt, ParaView, DV3D, ...)
- Workflow – data flows are directed graphs describing computational tasks
- Takes advantage of ESGF data management

ESGF Data Archive

ParaView Cluster parallel processing

Workflow

```

vslicer = load_workflow_as_function('vtdv3d.vt','slicer')
vslicer(variable='Relative_humidity')
vrrender = load_workflow_as_function('vtdv3d.vt','vr')
vrrender(variable='Relative_humidity')
    
```

Script

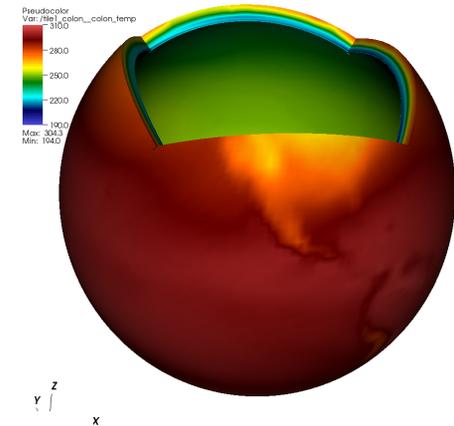
Provenance

9/5/12

D. N. Williams, LLNL Climate SFA Review

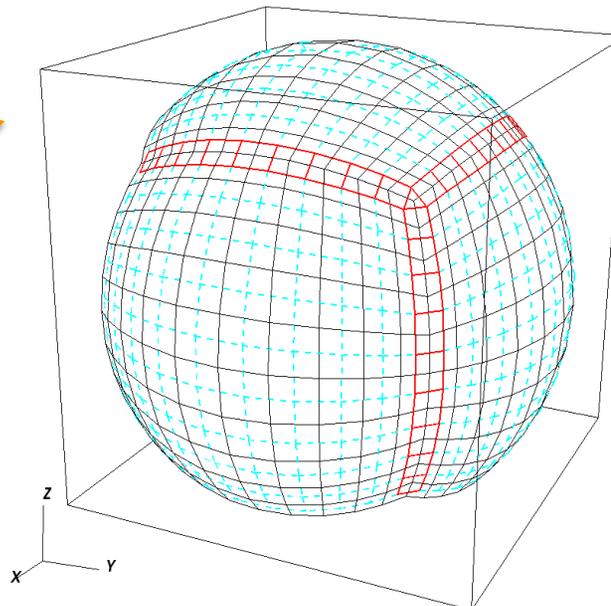
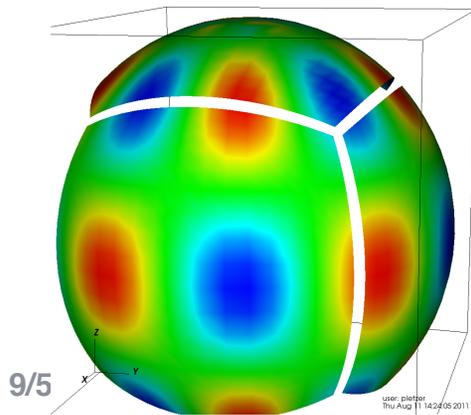
Support for Mosaic Grids: SCRIP, Gridspec, ESMF

- Supports cubed sphere & tripolar mosaic grids
- Complies with emerging Gridspec CF standard need for CMIP
- Recently added ESMF grids

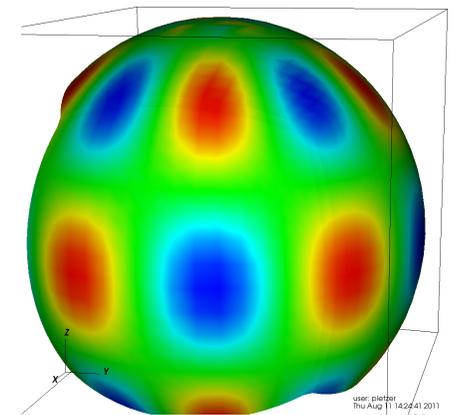


Atmospheric temperature from GFDL model

Create seam grids

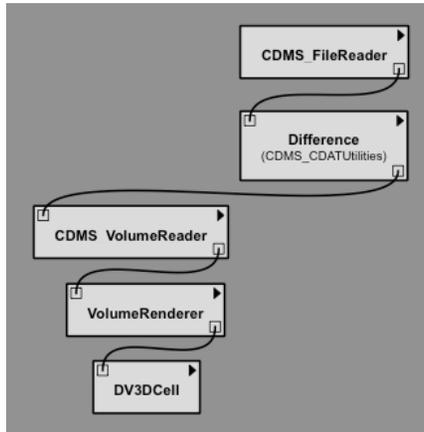


Fill gaps

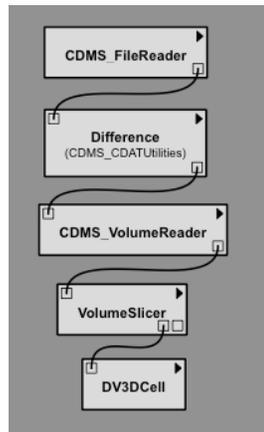


Exploitation of hardware and system parallelism for climate data analysis

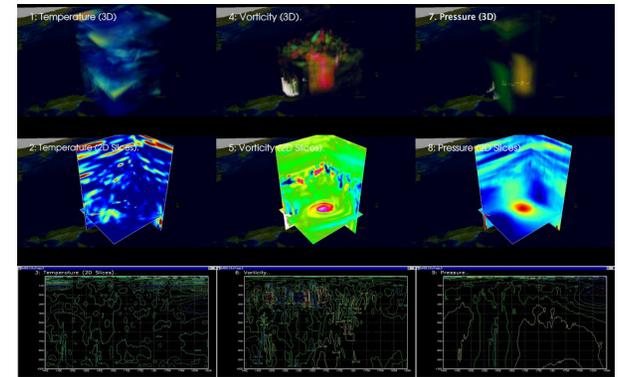
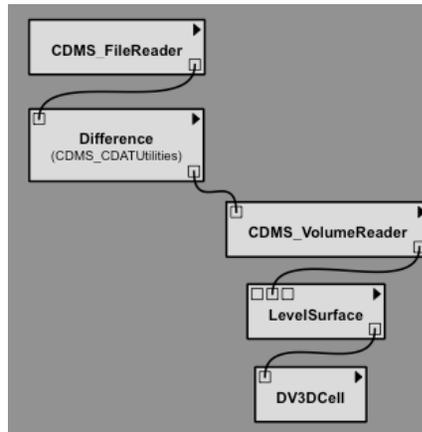
Client-0



Client-1

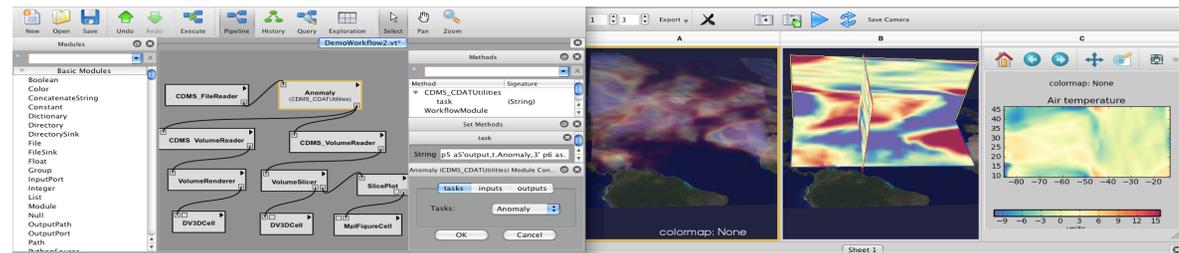
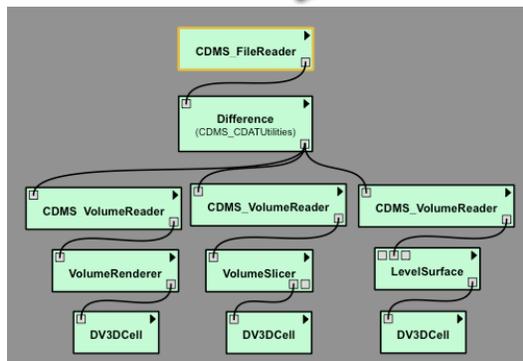


Client-2



- Data Parallelism
- Task Parallelism
- Time Parallelism

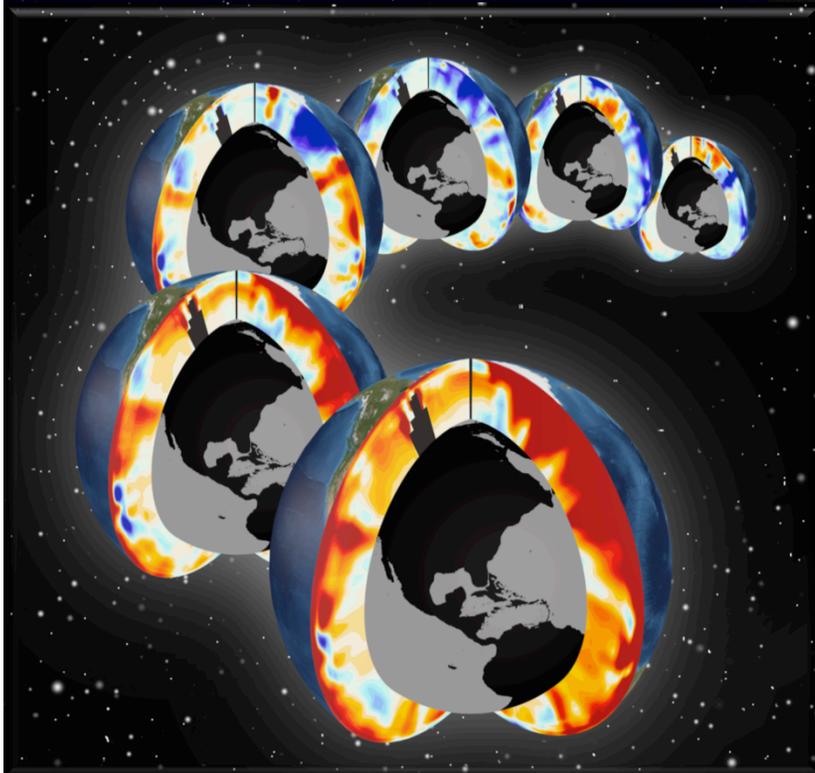
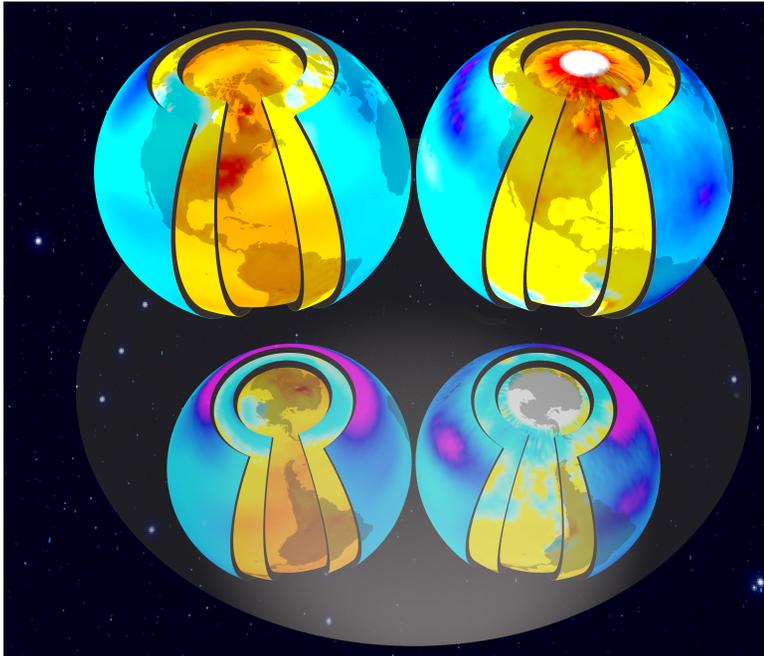
Server



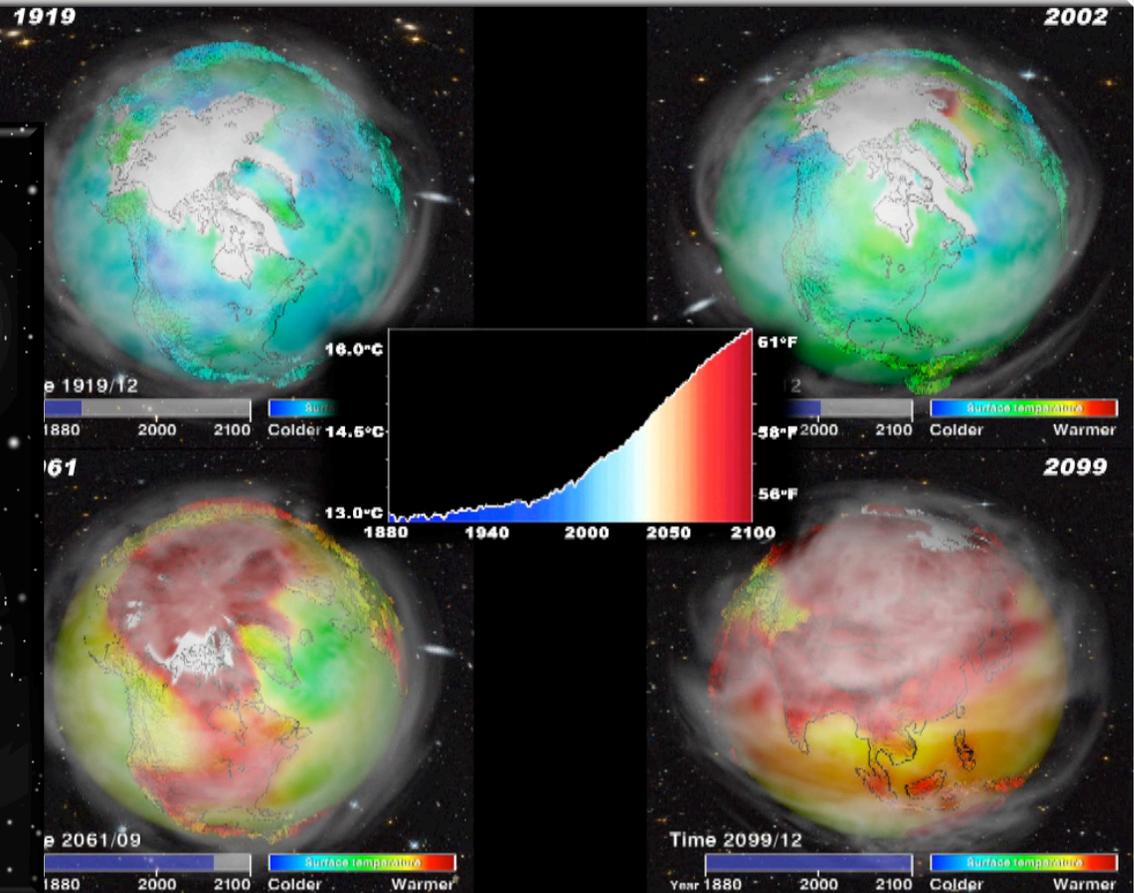
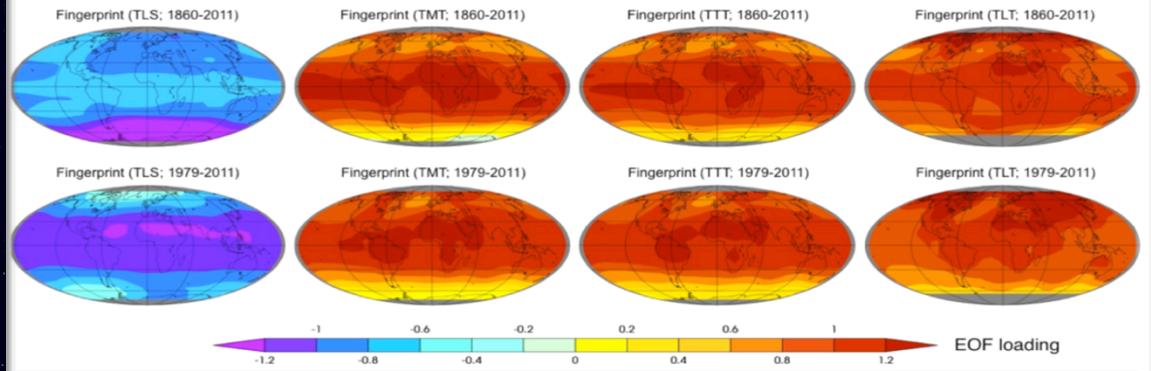
9/5/12

D. N. Williams, LLNL Climate SFA Review

Images and animations



Sensitivity of CMIP-5 O3+V Fingerprints to Length of Record

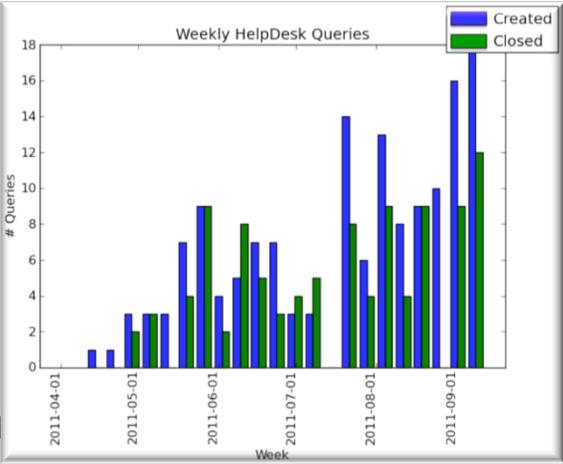


Important publications, conferences, and workshops



- Publication:** B.D. Santer, J. Painter, C. Mears, C. Doutriaux, P. Caldwell, J.M. Arblaster, P. Cameron-Smith, N.P. Gillett, P.J. Gleckler, J.R. Lanzante, J. Perlwitz, S. Solomon, P.A. Stott, K.E. Taylor, L. Terray, P.W. Thorne, M.F. Wehner, F.J. Wentz, T.M.L. Wigley, L. Wilcox, and C.-Z. Zou, “Identifying Human Influences on Atmospheric Temperature: Are Results Robust to Uncertainties?” Proceedings of the National Academy of Sciences
- Publication:** Jim Ahrens, Bruce Henderickson, Gabrielle Long, Steve Miller, Robert Ross, Dean Williams, “Data Intensive Science in the U.S. Department of Energy: Case Studies and Future Challenges,” IEEE Computer and Information Science and Engineering, volume PP issue 99, DOI 10.1109/MCSE.2011.77, 30 December 2011, (http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5999634&tag=1)
- Book:** ESM-Software book with ESGF contributions to the GRID Chapter 7, The Dean N. Williams et al. GRID chapter is special in that it is covering a very fluid subject, compared to many of the other chapters. First year college students are the intended audience of the book and it is expected to come out in early 2012. The publisher Springer will be offering smaller volumes of the book where each chapter would become its own mini-book. Please visit the following URL for more details: <http://www.springer.com/authors/book+authors/springerbriefs>
- Next Publication:** BAMS article, titled: “A Solution to Climate Science “
- ”:** Accessing Multi-Model Climate Simulation and Observation Data”, proposal abstract accepted, expected – end of 2012 or beginning of 2013

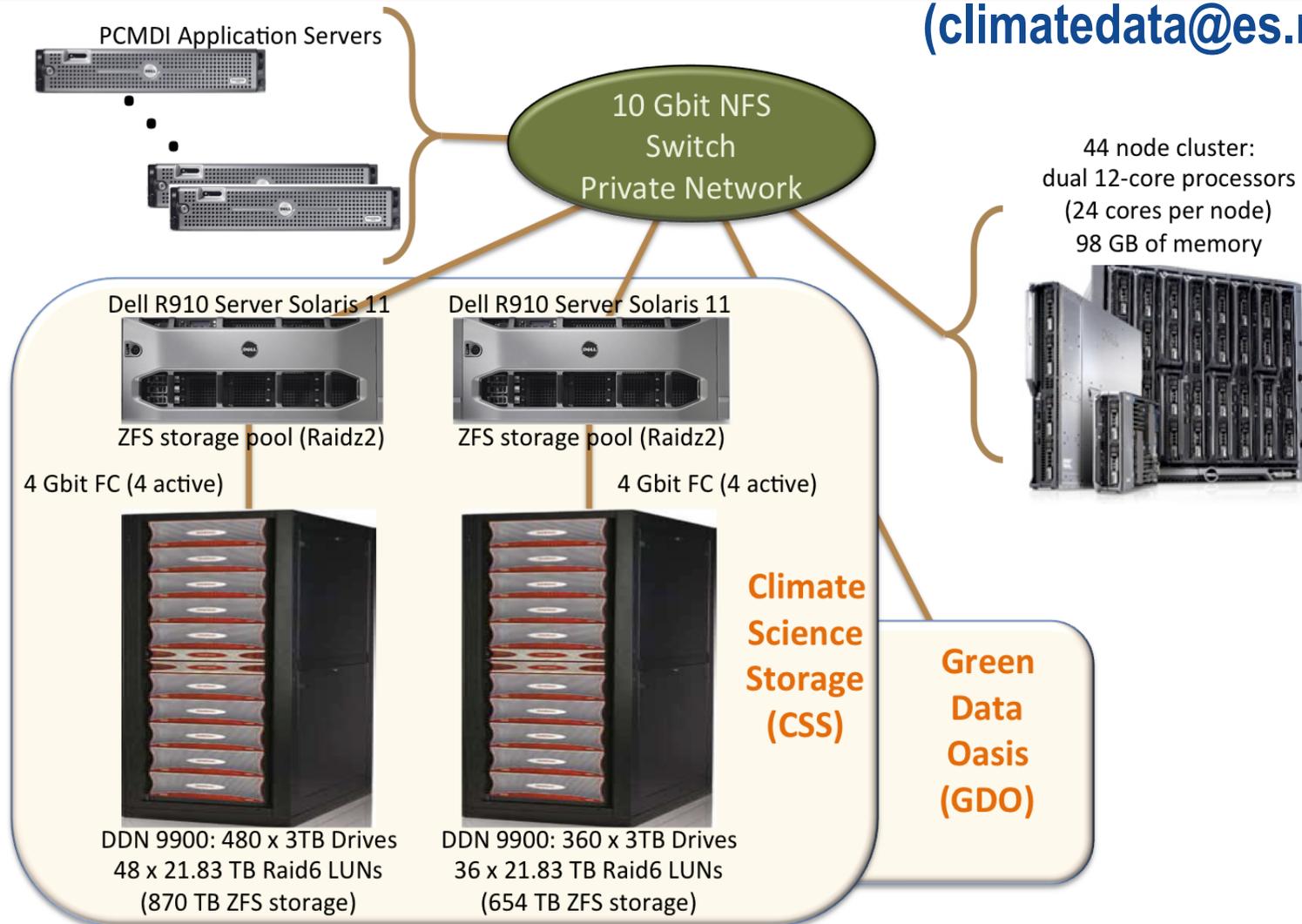
LLNL AIMS leads multiple collaborations



ASCR Climate 100 project: connecting our system to the largest network in the world and organizing the climate network community

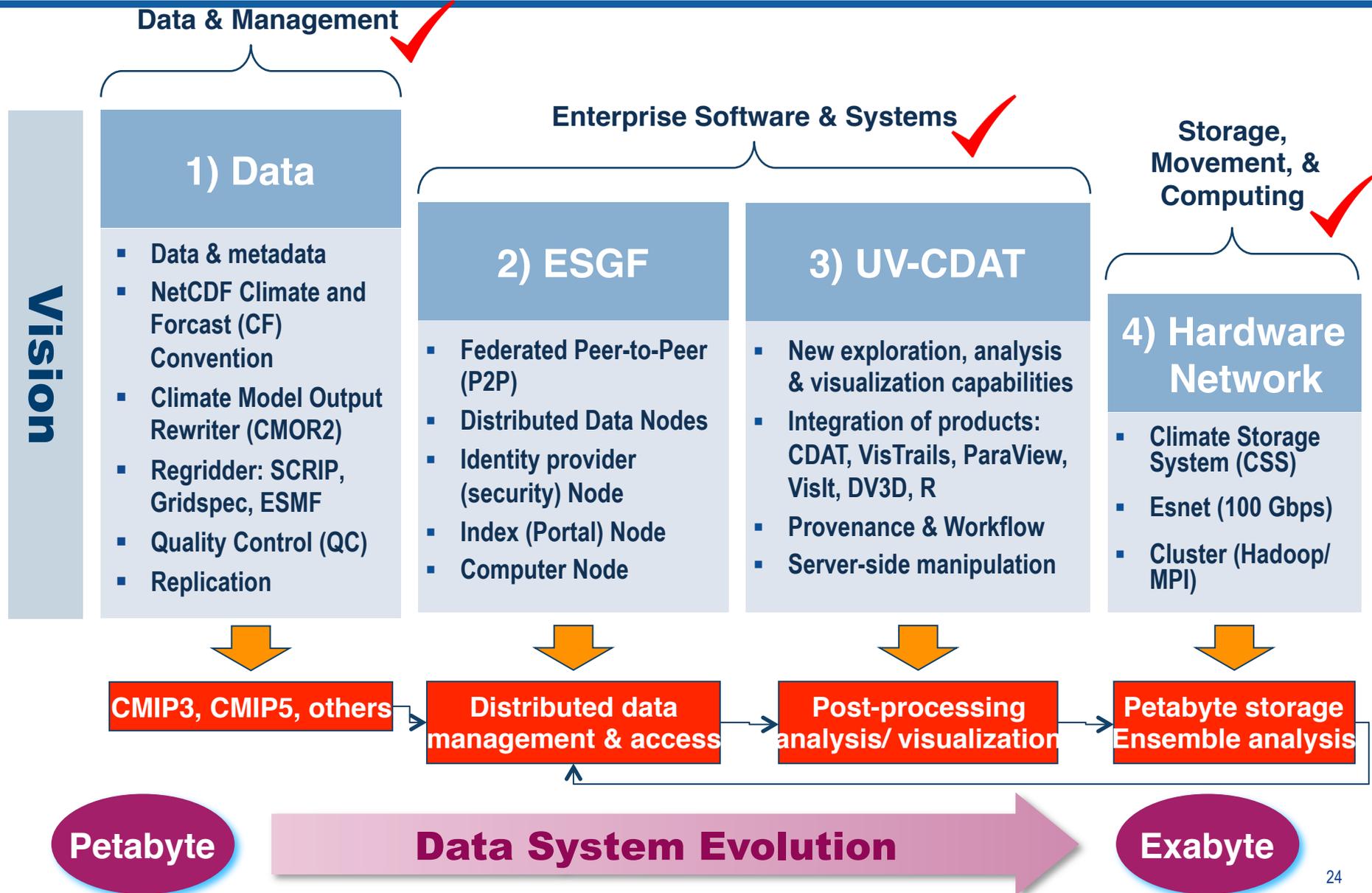
ESnet 100 Gbps Internet

(climatedata@es.net)



9/5/1:

Critical “vision” components, operations and services for climate change research productivity: leaders in climate “data science”



Why are we doing this? Complexity and size are on the rise

- In climate science, hundreds of exabytes are expected by 2020 *. Current and future heterogeneous climate data will be distributed around the globe and must be harnessed to find solutions to mission critical problems.
- More requirement; more constraints
 - Need to expand and integrate new modeling capabilities (e.g., prediction, uncertainty quantification (UQ), assimilation of more diverse data)
 - Staff expertise and competencies must be flexibly applied to multiple projects and programs
- Applications are getting more complex
 - Analysis of “state of the science” and future directions in areas of interest to climate research (e.g., test beds, UQ)
- Architectures are getting more complex
 - Heterogeneous applications
 - Data reduction to reduce data movement
 - Exploitation of new hardware and systems for greater science productivity

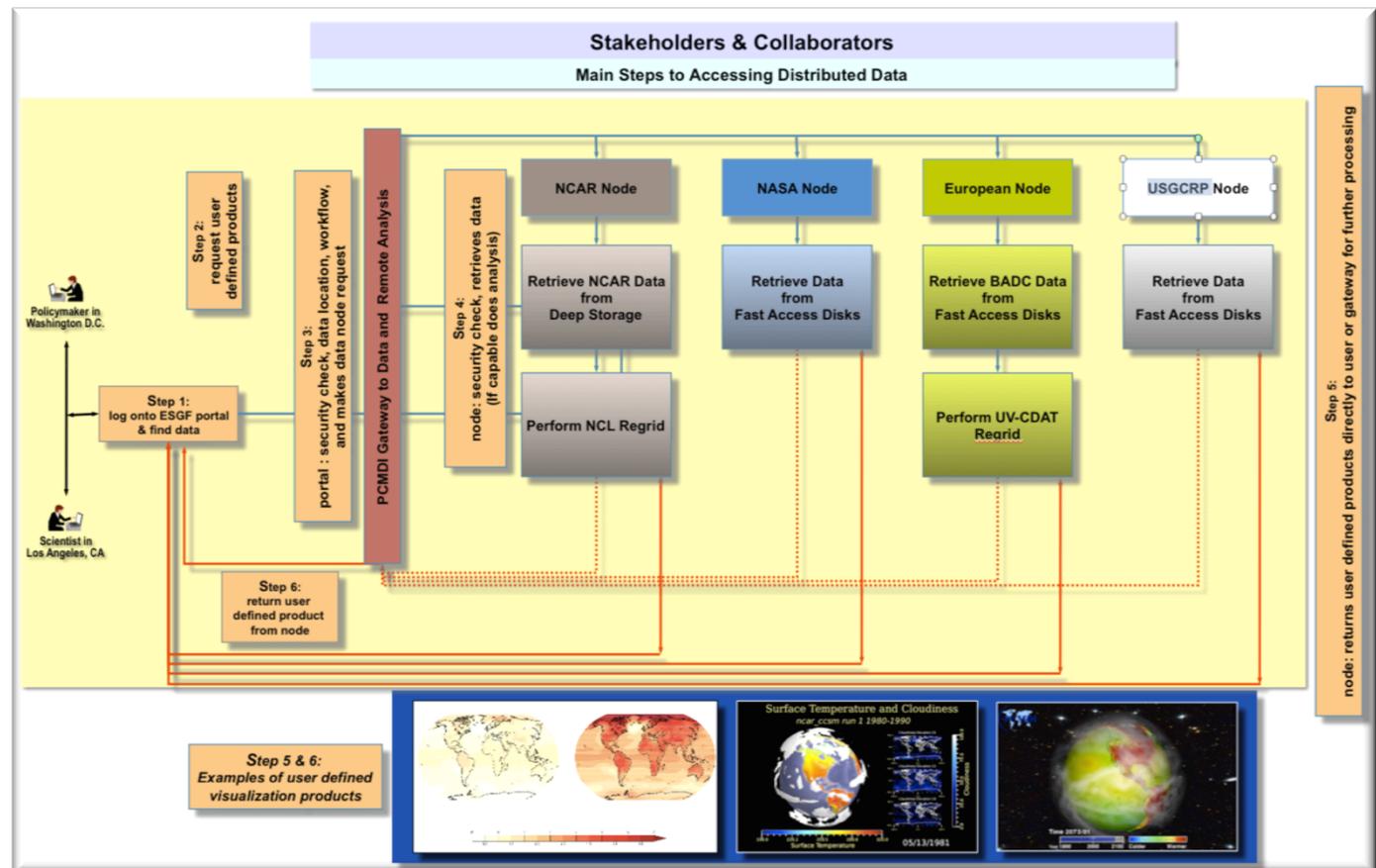


* 2011 Climate Knowledge Discovery Workshop, DKRZ, Hamburg, Germany

We need to understand the impact of these complexities and their interactions on application performance and science productivity.

Vision roadmap to future: next steps

- Continue to build the CMIP5 (IPCC AR5) data repository
- Refine the integration of UV-CDAT as a backend service to ESGF and interact directly with ESGF to allow full client-side usage
- Inclusion of observational and reanalysis data from NASA, NOAA, ORNL, others
- Community-driven open source software development and project governance



Funding of ongoing operational and software support for LLNL climate science data capabilities is a critical issue

- Software and Model Archive Federated Data System Components have been developed entirely as 3-5 year R&D projects through support from BER and ASCR.
- Over the last few years, we were generously given additional support by BER through a succession of one-time only funding actions, which are unlikely to continue.
- The current budget for AIMEs under the SFA provides only one FTE of support.
- Unlike the other BER/CESD data centers, we do not have a separate budget to fund ongoing infrastructure and support tasks.
- We estimate that 4-5 FTE's are needed:
 - For community outreach and engagement
 - participation metadata standards definition committees
 - Collaboration and infrastructure design and deployment with other BER and USGCRP climate data centers
 - Operational support of the ESGF-based CMIP model data distribution system
 - Ongoing ESGF software support
 - Assistance to modeling centers for software deployment and maintenance
 - Maintenance of CMIP replicated data archive housed at LLNL
 - Servicing requests driven by the IPCC, USGCRP, DOE and the research community
 -

LLNL stands ready to pursue research and provide software infrastructure support to meet DOE's research needs

- Our expertise is diverse
- Our record is strong
- Our commitment to excellence continues
- We know where we are going, but the vision is clear: to integrate independent projects to meet the strategic goals and lead the community in delivering a coherent Exascale scientific data management software system.

With your ongoing support and guidance, we will continue addressing climate change research issues of global importance.

LLNL's Analytics, Informatics, and Management Systems (AIMS)' Team (each team member has specific talents and tasks)

- Talents joined to work towards a common goal:
 - **Analytics**
 - Dan Bergmann
 - Charles Doutriaux
 - Dr. Elo Leung
 - Dr. Renata McCoy
 - Dr. Jeff Painter
 - many others in Livermore Computing
 - Eugenia Gabrielova (Student)
 - Dinorah C. Rodriguez (Student)
 - **Climate Science for a Sustainable Energy Future (CSSEF)**
 - Eddy Banks
 - Gavin Bell
 - **Data Operations**
 - Robert Drach
 - Michael Ganzberger
 - Dr. Renata McCoy
 - Dr. Jeff Painter
 - **Earth System Grid Federation (ESGF)**
 - Eddy Banks
 - Gavin Bell
 - Dr. Luca Cinquini
 - Robert Drach
 - Matthew B. Harris
 - Dr. Renata McCoy
 - **Hardware (storage, cluster, Networks, ...)**
 - Jenny Aquilino, Robin Goldstone, Tony Hoang, Jeff Long, Craig Schwonke, David Smith
 - **Ultra-scale Visualization Climate Data Analysis Tools (UV-CDAT)**
 - Dr. Timo Bremer
 - Charles Doutriaux
 - Robert Drach
 - Dr. Elo Leung
 - Dr. Jeff Painter
 - **Web Applications and UI development**
 - Charles Doutriaux
 - Dr. Elo Leung
 - Dr. Renata McCoy
 - **Visualization**
 - Dr. Timo Bremer
 - Charles Doutriaux
 - Rich Cook, Eric Brugger, Ross Gaunt, and others from Livermore Computing
 - **Dean N. Williams (Principal Investigator: AIMS, ESGF, UV-CDAT; co-PI: CSSEF, GO-ESSP, Data Exploration, GIP, Climate 100) - Project Lead**

Questions and discussion

