4TH ANNUAL

EARTH SYSTEM GRID FEDERATION AND ULTRASCALE VISUALIZATION CLIMATE DATA ANALYSIS TOOLS

FACE-TO-FACE CONFERENCE

DECEMBER 2014

A GLOBAL CONSORTIUM OF GOVERNMENT AGENCIES, EDUCATIONAL INSTITUTIONS, AND COMPANIES DEDICATED TO DELIVERING ROBUST DISTRIBUTED DATA, COMPUTING LIBRARIES, APPLICATIONS, AND COMPUTATIONAL PLATFORMS FOR THE NOVEL EXAMINATION OF EXTREAME-SCALE SCIENTIFIC DATA.

Day 1: Tuesday, 9 December 2014
User Feedback and Project Requirements

Session Two: ESGF Governance

Title and Presenter	Abstract
ESGF Governance ***Justin Hnilo (DOE BER CESD, Justin.Hnilo@science.doe.gov)***	Although initiated by DOE, the Earth System Grid Federation (ESGF) has become a multi-agency, international collaboration critical to the success of archiving, delivering, and analyzing well-known climate data, including the Working Group on Coupled Modeling's Coupled Model Intercomparison Project (CMIP) data used for the Intergovernmental Panel on Climate Change assessment reports. Now that the ESGF infrastructure has been adopted by many other projects and supporting CMIP activities, there is a need to establish a more formal governance structure that ensures resource implications for users of disparate data as well as funding agencies. The presentation and discussion will touch upon the status of the ESGF governance in respect to the current funding agencies: Department of Energy, EU Commissions, National Aeronautics and Space Administration, National Oceanic and Atmospheric Administration, and the Australian Department of Education.

Session Three: Project Feedback and Requirements

Title and Presenter	Abstract
Significance of the User Support Process in ESGF **(Session's Keynote)** ***Hashim Chunpir (DKRZ Hamburg, chunpir@dkrz.de)***	The Earth System Grid Federation (ESGF) infrastructure is a collection of technologies, systems, people, policies, practices, processes, and relationships that interact with each other in a federated environment. The ESGF infrastructure came into being as a result of a need, and the need was to fulfill data-driven climate community research. As the ESGF reached its production level and became a commodity, it started serving diverse user communities, particularly from the domain of Climate Science. The key to envisioning a usable and a healthy e-infrastructure today is to streamline the user support process and to try to address the communication and standardization challenges that arise between the stakeholders of the ESGF. This talk will briefly describe the current user support scenario for ESGF based on the empirical findings collected in the past and highlight the challenges that we face today that must be addressed in future. Addressing user and usability issues within the ESGF infrastructure will not only attract the user communities but also will provide user satisfaction, cost efficiency, sustainability of e-infrastructure, and promote innovation and development of the infrastructure, resulting in a boom in the data-driven e-research.
WGCM Infrastructure Panel Requirement ***Karl Taylor (DOE/LLNL, taylor13@llnl.gov)***	The newly formed WGCM Infrastructure Panel (WIP) is charged with attempting to articulate and set priorities for what is needed in the way of climate modeling infrastructure from the perspective of the modeling groups and the WGCM. The goal is to establish standards and policies for sharing climate model output that ensure consistency across WGCM activities. The WIP is currently developing four white papers describing key aspects of the infrastructure needed to support Coupled Model Intercomparison Project Phase 6, and these will be discussed with a focus on those components of particular interest to the ESGF community.
CMIP6 and MIPs ***Karl Taylor (DOE/LLNL, taylor13@llnl.gov)***	The Coupled Model Intercomparison Project has been restructured to enhance its scientific impact while reducing the burden placed on modeling centers and the modeling infrastructure. CMIP now calls for an ongoing small set of benchmark "Diagnosis, Evaluation, and Characterization of Klime Experiments" (DECK) that will be performed as part of the model development cycle at each modeling centers. Building on these, Coupled Model Intercomparison Project Phase 6 (CMIP6) will offer a smorgasbord of additional MIPs that address specific science questions. Modeling groups will be free to tailor their participation in CMIP6 according to their resource limits and scientific interests. In the context of the new CMIP6 design, implications for modeling infrastructure will be discussed.

High–End Computing Program Manager and Weather Focus Area Program Scientist *Tsendgar Lee (NASA HQ, tsengdar.j.lee@nasa.gov)*	Earth system scientists are facing significant data analysis challenges as observation and model-output data become bigger, which is caused by the spatial, temporal, and spectral resolutions becoming higher and the data records becoming longer. The information systems developed and architected in the past will need to be upgraded and enhanced to face the big data challenges. In this talk we will discuss how the space agencies are confronting the challenges by putting together a next-generation system architecture that couples data, compute, storage, and tools. An open-source strategy has been established to enable open development and open collaboration.
Collaborative REAnalysis Technical Environment-Intercomparison Project (CREATE-IP) and ana4MIPs: Enhancing Access to Reanalysis Using ESGF *Jerry Potter (NASA/GSFC, gerald.potter@nasa.gov)*	NASA/GSFC is gathering gridded reanalysis data from major weather forecast centers around the world and saving them side-by-side on the Earth System Grid Federation (ESGF) to help better understand the patterns responsible for such phenomena as heat waves, droughts, and floods and ultimately improve climate model predictions. The CREATE-IP, or Collaborative REAnalysis Technical Environment-Intercomparison Project, has created a repository of reanalyses (essentially re-forecasts of past weather using the latest forecast models) that can help improve weather and climate forecasts by studying the differences and similarities among various reanalysis efforts. Participating organizations include the National Aeronautics and Space Administration (NASA), the National Oceanic and Atmospheric Administration's National Centers for Environmental Prediction and Earth System Research Laboratory, the European Centre for Medium-Range Weather Forecasts, and the Japan Meteorological Agency. In addition to distributing the data on ESGF, NASA's Climate Model Data Services group is partnering with the NASA Center for Climate Simulation to develop tools that will bring to bear the massive computing power of the latest supercomputers along with large data storage facilities to make analysis and comparison of complex model output faster and more efficient for climate scientists.
ESGF Functionality Needed for obs4MIPs and Other Data Sets *Robert Ferraro (NASA/JPL, robert.d.ferraro@jpl.nasa.gov)*	Obs4MIPs is delivering satellite observations formatted in the same manner as the Coupled Model Intercomparison 5 model outputs via the Earth System Grid Federation (ESGF). But there are differences between the observations' attributes (metadata) and the model attributes that do not mesh well with either the DRS or the standard search capabilities. And the agencies that support these data sets are asking for reporting of usage statistics. As obs4MIPs grows, and other projects like ana4MIPs are added, some evolution of the ESGF search and data delivery capabilities will be needed to support a more diverse user base and a wider variety of file types.
CORDEX *Sébastien Denvil (IPSL/IS-ENES2, sebastien.denvil@ipsl.jussieu.fr)*	Global climate models are at the basis of climate change science and of the provision of information to decision-makers and a large range of users. Within Europe, the European Network for Earth System Modeling (ENES) gathers together the European climate/Earth system modeling community, which is working on understanding and prediction of future climate change. ENES, through IS-ENES (phase 1 and 2), promotes the development of a **common distributed modeling research infrastructure** in Europe in order to facilitate the development and exploitation of climate models and better fulfill the societal needs with regards to climate change issues. IS-ENES2 gathers 18 partners from 10 European countries and includes the 6 main European Global Climate Models. IS-ENES combines expertise in climate **Earth system modeling (ESM)**, in **computational science**, and in studies of **climate change impacts**. This talk will highlight efforts undertaken under the IS-ENES2 coordination to support the dissemination of CORDEX on the ESGF. By many aspects this efforts prefigure what will be done for the Coupled Model Intercomparison Project Phase 6.
The climate4impact Portal: Bridging the CMIP5 and CORDEX Data Infrastructure to Impact Users *Maarten Plieger (KNMI/IS-ENES, plieger@knmi.nl), Wim Som de Cerff, Christian Page, Natalia Tatarinova, Ronald Hutjes, Fokke de Jong, Lars Barring, and Elin Sjökvist*	The aim of climate4impact is to enhance the use of climate research data and to enhance the interaction with climate effect/impact communities. The portal is based on 17 impact use cases from five different European countries and is evaluated by a user panel consisting of use case owners. It has been developed within the European projects IS-ENES and IS-ENES2 for more than five years, and its development currently continues within IS-ENES2. The focus is mainly on the scientific impact community due to the community's breadth. This work has resulted in the ENES portal interface for climate impact communities and can be visited at www.climate4impact.eu. The climate4impact is connected to the Earth System Grid Federation (ESGF) nodes containing global climate model data (GCM data) from the fifth phase of the Coupled Model Intercomparison Project (CMIP5) and regional climate model data (RCM) data from the Coordinated Regional Climate Downscaling Experiment. This global network of climate model data centers offers services for data description, discovery, and download. The climate4impact portal connects to these services using OpenID, and offers a user interface for searching, visualizing, and downloading global climate model data and more. A challenging task was to describe the available model data and how it can be used. The portal tries to inform users about possible caveats when using climate model data. All impact use cases are described in the documentation section, using

	highlighted keywords pointing to detailed information in the glossary. During the project, the content management system Drupal was used to enable partners to contribute on the documentation section. In this presentation, the architecture and following items will be detailed: • Visualization: Visualize data from ESGF data nodes using ADAGUC Web Map Services. • Processing: Transform data, subset, export into other formats, and perform climate indices calculations using Web Processing Services implemented by PyWPS, based on NCAR NCPP OpenClimateGIS and IS-ENES2 icclim. • Security: Login using OpenID for access to the ESGF data nodes. The ESGF works in conjunction with several external websites and systems. The climate4impact portal uses X509-based short-lived credentials, generated on behalf of the user with a MyProxy service. Single Sign-on (SSO) is used to make these websites and systems work together. • Discovery: Faceted search based on e.g. variable name, model, and institute using the ESGF search services. A catalog browser allows for browsing through CMIP5 and any other climate model data catalogues (e.g. ESSENCE, EOBS, UNIDATA). • Download: Directly from ESGF nodes and other THREDDS catalogs. This architecture will also be used for the future Copernicus platform, developed in the EU FP7 CLIPC project. • Connection with the downscaling portal of the university of Cantabria. • Experiences on the question and answer site via Askbot. There are two current main objectives for climate4impact. The first one is to work on a web interface, which automatically generates a graphical user interface on WPS endpoints. The WPS calculates climate indices and subset data using OpenClimateGIS/icclim on data stored in ESGF data nodes. Data is then transmitted from ESGF nodes over secured OpenDAP and becomes available in a new, per user, secured OpenDAP server. The results can then be visualized again using ADAGUC WMS. Dedicated wizards for processing of climate indices will be developed in close collaboration with users. The second one is to expose climate4impact services, so as to offer standardized services which can be used by other portals. This has the advantage of adding interoperability between several portals, as well as enabling the design of specific portals aimed at different impact communities, either thematic or national.
Learning About Data Systems and Data Tools from Practitioner Applications **(Set up go to meeting)** ***Ricky Rood (University of Michigan, rbrood@umich.edu)***	As discussed in Rood and Edwards (2014; http://earthzine.org/2014/05/22/climate-informatics-human-experts-and-the-end-to-end-system/) there are substantial barriers to the usability of climate data by both scientific and non-scientific users. The conclusions in Rood and Edwards were drawn from several sources: experience in the Great Lakes Integrated Sciences and Assessments project, the National Climatic Predictions and Projections Platform, academic literature on usability, and formal evaluations by students in a University of Michigan class, Climate Informatics. A fundamental misconception remains in many data systems and data services in the climate community; namely, that providing accessibility of data is adequate to assure the broad use of data. The perpetuation of this loading-dock model of data provision assures that data and knowledge are used by those most vested in the communities, e.g. climate scientists and those specifically funded in projects that require the use of the data. To move data off the loading dock requires services that understand the use cases of the end users and contribute to a chain of tools and services, including training on what to do with the data and how to do it. In numerous interviews with end users, basic barriers such as glossaries that define scientific terms as well as arcane files names were cited as barriers that motivated users to abandon online services in a matter of minutes. Other barriers included data formats unknown in user communities and difficulty in developing interfaces with a community's tools. These barriers in concert with broader issues of tailoring data and knowledge to specific applications led to the need for data systems to support the roles of human intercessory in the chain connecting data providers with data users.
Model Output Evaluation and Data Dissemination for Seasonal and Shorter Time Scales: NMME and HIWPP ***Cecelia De Luca (NOAA/ESRL, cecelia.deluca@noaa.gov) and Eric Nienhouse (NSF/NCAR, ejn@ucar.edu)***	The evaluation of model outputs for seasonal and shorter term predictions is being coordinated through projects that involve efforts from multiple modeling and forecast centers. These coordinated efforts can introduce high-volume data products and new challenges to distributed data systems such as the Earth System Grid Federation (ESGF). In this talk we describe two such efforts, the National Multi-Model Ensemble and the High Impact Weather Prediction Project. We'll present the projects' use of ESGF and the CoG user interface as a project documentation, information dissemination, data discovery, and data access infrastructure.

The ACME Modeling Project Infrastructure ***Dave Bader (DOE/LLNL, ACME Project Council Chair, bader2@llnl.gov)***	The Accelerated Climate Modeling for Energy (ACME) project is a newly launched project sponsored by the Earth System Modeling program within DOE's Office of Biological and Environmental Research. ACME is an unprecedented collaboration among eight national laboratories, the National Center for Atmospheric Research, four academic institutions, and one private-sector company to develop and apply the most complete, leading-edge climate and Earth system models to the most challenging and demanding climate-change research imperatives. It is the only major national modeling project designed to address U.S. Department of Energy (DOE) mission needs and efficiently use DOE Leadership Computing resources now and in the future. ACME will achieve its goals through four intersecting project elements: 1. A series of **prediction and simulation experiments** addressing scientific questions and mission needs; 2. A well-documented and tested, continuously advancing, evolving, and improving **system of model codes that comprise the ACME Earth system model**; 3. The ability to use effectively **leading (and "bleeding") edge computational facilities** soon after their deployment at DOE national laboratories; and 4. **An infrastructure** to support code development, hypothesis testing, simulation execution, and analysis of results. This talk will describe how the Ultrascale Visualization Climate Data Analysis Tools and the Earth System Grid Federation are essential pieces for the development of the scalable infrastructure to enable science simulation at the exascale. The priority science drivers and resulting three-year experiments were used to define the functionality of the initial simulation system. Initial infrastructure design was based on the requirements to facilitate hypothesis-testing workflows (configuration, simulation, diagnostics, and analysis). The infrastructure element will be continuously evolving. It will maintain a disciplined software engineering structure and develop turnkey workflows to enable efficient code development, testing, simulation design, experiment execution, analysis of output, and distribution of results within and outside the project.
A User's Perspective on Acquisition and Management of CMIP5 Data ***Jennifer Adams (COLA/NOAA, jma@cola.iges.org)***	The complexity, volume, and distributed nature of the Coupled Model Intercomparison Project Phase 5 (CMIP5) data collection has left many users struggling to acquire the CMIP5 data they need. This presentation outlines strategies that were developed to overcome the challenges CMIP5 data users face: authentication, searching for published data that match a list of desired experiments and variables, acquisition of wget scripts, managing wget script execution and the high wget failure rate, retention of critical metadata not present in the data files, version control, local data management, and setting up the data for analysis and visualization using GrADS. All these strategies exist in an automated workflow that is completely independent of any browser interface.
The GeoMIP Perspective on Interactions with ESGF **(set up go to meeting)** ***Ben Kravitz (Pacific Northwest National Laboratory, ben.kravitz@pnnl.gov)***	In this talk, I discuss some of the strengths and weaknesses of the Earth System Grid Federation (ESGF) as I have seen through my work on the Geo-engineering Model Intercomparison Project (GeoMIP). ESGF has provided an excellent common framework so that all of the climate model output necessary for conducting GeoMIP analysis is available in one place in a standard format; such coordinated efforts are necessary for projects of the magnitude of the Coupled Model Intercomparison Project Phase 6. However, the issues we have encountered in both hosting and retrieving climate model output are substantial. Establishing a data node has proven to be costly and time consuming, and transferring output to other nodes for hosting has met with moderate success. Downloading climate model output from ESGF is most reliably done one file at a time due to complications with certificate authentication. Although some nodes are well set up for wget or Globus Online transfer, we have yet to discover a universal method for downloading large numbers of files from ESGF.

Session Four: Modeling and Data Center Requirements

Title and Presenter	Abstract
Modeling and Data Center Requirements **(Session's Keynote)** ***Sébastien Denvil (IPSL/IS-***	Earth system model simulations are central to the study of complex mechanisms and feedbacks in the climate system and to provide estimates of future and past climate changes. Recent trends in climate modeling are to add more physical components in the modeled system, increasing the resolution of each individual component and the more systematic use of large suites of simulations to address many scientific questions. Climate simulations may therefore differ in their initial state, parameter values, representation of physical processes, spatial resolution, model complexity, and

ENES2, sebastien.denvil@ipsl.jussieu.fr)	degree of realism or degree of idealization. In addition, there is a strong need for evaluating, improving, and monitoring the performance of climate models using a large ensemble of diagnostics and better integration of model outputs and observational data. At the same time, the Data and Supercomputing Center offers services to several communities and to specific communities (like the climate modeling community). This talk will try to gather common general requirements that must be fulfilled to ensure the Center's acceptance of a system such as the Earth System Grid Federation (ESGF) and to ensure its willingness to support and contribute to ESGF.
Australia (ANU/NCI) ***Ben Evans (ANU/NCI, Ben.Evans@anu.edu.au)***	NCI provides a high performance collaborative center for the Australian research community that spans national science agencies and research institutions. A particular initiative has been to provide a high-performance data and high-performance computing environment that is suited for both modeling and analyzing the whole earth system. To achieve this at a high standard for both research and government outcomes, there have been significant improvements to the breadth of services and their functionality, their integration with the underlying hardware (supercomputing, cloud, visualization, and data storage), the overall manageability and provenance management, and flexibility for ongoing expansion. The core of this includes the data management (both metadata and data), the tools (both standard and collaborative development), data services, and community environments (virtual laboratories). The Earth System Grid Federation (ESGF) infrastructure is a significant component of our infrastructure, and a flagship for international collaborative infrastructure, but it is just one part in this dizzying array of components. I will highlight how the ESGF currently fits for us and consider the challenges of infrastructure that is highly connected, relevant to research needs, nationally and internationally trusted, aligned with our world peers, and stays on the critical path to help meet the future needs.
IPCC-DDC (DKRZ) ***Martina Stockhause (WDCC/DKRZ, stockhause@dkrz.de)***	DKRZ hosts the IPCC Data Distribution Center, which provides long-term access to Coupled Model Intercomparison Project data for interdisciplinary (re-)use. Beyond permanent and persistent data access, the Center must provide detailed documentations, a uniform data quality, and DataCite DOI data citations to enable data users to accept or even trust the data and to give credit to data creators.
France (IPSL) ***Sébastien Denvil (IPSL/IS-ENES2, sebastien.denvil@ipsl.jussieu.fr)***	The Earth System Grid Federation (ESGF) has been operational in France since April 2011. There are six institutions running ESGF in France. Three of them are modeling groups—IPSL, CNRM, and CERFACS—and three of them are National Supercomputing and Data Centers—TGCC, IDRIS, and CINES. This talk will present how the French partners organized themselves to ensure smooth operations and effective contributions to the ESGF federation.
The Status of ESG-BNU Node in China ***Baogang Zhang, China (Beijing Normal University)***	The Earth System Grid Federation (ESGF) is an operational system for serving climate data from multiple locations and sources. The ESG-BNU node, established in 2012, is one of 5 ESGF nodes in China. It is also the only IdP/index node in China. At present, the ESG-BNU node has published 17 experiments of 10,595 GB data set for CMIP5 and 4 experiments of 1,174 GB data set for GeoMIP—both generated by the Earth System Model of Beijing Normal University (BNU-ESM). It has provided over 31.02 TB of data with an average 700 KB/s download speed to scientists all over the world. As an IdP/index node, we try to make replicas of other model centers so that the Chinese data user can access the data sets much faster. We find that over 95% of data requests are from China, U.S. and Japan. About 75% downloads come from historical, rcp45 and rcp85 experiments and 79% of downloads come from the monthly data sets. As for the upcoming CMIP6, data volume of one model center will reach to over 100 TB. Making replica of all experiments maybe too expensive and not necessary for many ESGF nodes. We hope such download statistical analysis can be helpful for ESGF nodes to decide which experiments should be replicated or be replicated with priority.
Experiences with the ESGF Data Portal at CCCma **(Set up Go-To-Meeting)** ***Slava Kharin (Canadian Centre for Climate Modelling and Analysis, Slava.Kharin@ec.gc.ca)***	We present some experiences at CCCma dealing with the ESGF data portal during the CMIP5 exercise and suggest a few areas for improvements.
UK (BADC) ***Phil Kershaw (BADC/IS-ENES2, philip.kershaw@stfc.ac.uk)***	The BADC is one of four data centers operated by CEDA, The Centre for Environmental Data Archival on behalf of the UK Natural Environment Research Council. CEDA receives it overarching requirements for engaging ESGF through NERC: to maximize the UK's contributions to the CMIP cycle and exploitation of the data for the user communities it serves. Alongside this, there are number of supplementary requirements related to CEDA's stakeholders to which it is contracted to provide services. These include the UK Met Office, UK government, IPCC,

	European climate community and others. Over the past years, we have seen international collaboration as a key to meeting these objectives: engaging with shared software development effort was more likely to result in systems fit for purpose and builds a community upon which to pool resources to create common tools and services. With the growth of ESGF as an operational system, however there is a need to address a number of additional critical factors: how to best integrate ESGF services with the rest of CEDA's evolving infrastructure, support for multiple projects within the federation and the ongoing operation and maintenance of the system and expected levels of service providers can meet. In this presentation we will explore these challenges.
NSF NCAR Data Center Requirements: Reducing Barriers to Community Data Products (USA) ***Eric Neinhouse (NSF/NCAR)***	As a national center NSF-NCAR manages over 5PB of climate and related data products. Integration of these products with distribution systems such as ESGF removes barriers to scientific data discovery and access. In this talk we will present key requirements and challenges for including these valuable data products in ESGF: • Publication of existing data products • Challenges of data discovery and use • The power of use metrics from distributed systems • Integration with tertiary storage systems • The need for software and content governance
GFDL Model Data Requirements and ESGF (USA) ***Serguei Nikonov* (NOAA/GFDL,** *Serguei.Nikonov@noaa.gov)*	The last two Intergovernmental Panel on Climate Change reports drove the architecture of the Geophysical Fluid Dynamics Laboratory (GFDL) data portal and brought it to its current state. During the Coupled Model Intercomparison Project Phase 5 project, GFDL was running two Data Portal infrastructures—ESGF and GFDL Curator. Running both gave us a good opportunity to compare strong and weak sides of both systems and elaborate requirements to the "ideal" system. For example, a useful feature would be interchangeable modularity of Earth System Grid Federation (ESGF) architecture for possibility to incorporate local model center existing subsystems element into ESGF. Some examples of such GFDL infrastructure elements will be discussed for consideration.

Session Five: Network Requirements

Title and Presenter	Abstract
Network Requirements—ICNWG **(Session's Keynote)** ***Eli Dart (DOE/ESnet, dart@es.net)***	The International Climate Network Working Group has been concentrating on improving data transfer performance for replication activities between five sites: ANU/NCI, BADC, DKRZ, KNMI, and PCMDI/LLNL. This talk will provide an update on the progress of the ICNWG efforts, and will discuss possible future scenarios for high-speed data replication and transfer between major data centers and computing facilities. If time permits, some highlights from the recent Climate CrossConnects meeting in Boulder, CO, will be discussed as well.

Day 2: Wednesday, 10 December 2014
ESGF and UV-CDAT Technical Presentations and Discussions

Session Six: ESGF Technical Development

Title and Presenter	Abstract
Technical Developments for the Community ***Dean N. Williams (DOE/LLNL, williams13@llnl.gov)***	The community of technical team members, consisting of computational and climate scientists, worked across institutional boundaries to develop and integrate software packages and sub-components for facilitating climate research. This effort enabled scientists in their daily activities and aided in the publication of hundreds of scientific articles. The developed partnership among the sponsors, institutions, universities, and private companies used the best human–computer interactions theory–based approach, coupled with pragmatic implementation of the system–user interface and modes of interaction. The scaled Agile development practice took full effect, highlighting individual roles, teams, and activities. This allowed team members to adjust schedules and priorities as necessary to quickly provide new solutions and meet the community's ever-changing needs. The goal of this presentation is to show integration of the many software components and how they relate to existing and future community projects.

ESGF Installation Working Team ***Nicolas Carenton (IS-ENES/IPSL, nicolas.carenton@ipsl.jussieu.fr) and Prashanth Dwarakanath (Linköping University, pchengi@nsc.liu.se)***	The Earth System Grid Federation (ESGF) installation working team was created in March 2014. Its main responsibilities are ESGF releases management, installation tools maintenance as well as node administrators support. One of its most important challenges is to provide an automated installation of a node that can complete in less than an hour. We will present here the work done since the team creation which includes the recovery of the ESGF build process, the implementation of several distribution mirrors, the improvement of the release management process as well as the new test and validation tools. We will also present the major releases of the year and the upcoming work on the installer that will lead to easier installation for administrators.
CoG: The New ESGF User Interface ***Luca Cinquini (NOAA/ESRL, Luca.Cinquini@jpl.nasa.gov)***	The Earth System CoG is a web interface that organizes data distribution for a multitude of projects in a federated and distributed environment. CoG will soon be replacing the current ESGF web user interface (i.e. the "web-front-end" module). Over the past year, the CoG team has worked through the tasks that needed to be accomplished to make the ESGF/CoG merging possible. Major upgrades in CoG functionality for the ESGF user community include a more powerful and flexible search interface, a model for exchanging information among peer nodes, streamlined group registration, and co-location of project data and documentation, just to name a few. This talk will review the status of the CoG development for ESGF adoption, and the last steps needed to execute the switch.
Publication as a Service: Globus Publish to ESGF ***Sasha Ames (DOE/LLNL) and Rachana Ananthakrishnan (U. of Chicago, ranantha@uchicago.edu)***	This talk will describe our progress made in the design and implementation of "publication-as-a-service" for the Earth System Grid Federation (ESGF). Publication-as-a-service takes the "system administration" overhead of data publication out of the hand of scientists, whose goal is to make their data available to the community. We will present the web interface to the service and details the backend processes that show the interaction of the web service with the ESGF publisher infrastructure. Additionally, we will discuss some recent changes to the publisher software that fit with the goals for supporting publication as a service.
Quality Control Working Team: esgf-qcwt ***Martina Stockhause (IS-ENES2/DKRZ, stockhause@dkrz.de) and Katharina Berger (IS-ENES2/DKRZ)***	Within the quality team, we consider all questions of how the Earth System Grid Federation (ESGF) could be improved in order to increase the quality of ESGF data services. The main working task is the integration of external information into ESGF, such as information on provenance, quality, and data citation. This implies the storage of data unpublishes events for provenance and the support of data collections (granularity of data citations). Requirements out of the WGCM Infrastructure Panel for the Coupled Model Intercomparison Project Phase 6 for this team will be integrated. Close collaborations with the ESGF teams on "Replication and Versioning" and "Publication" are required. We plan to give a demonstration to show the first results of the team.
ESGF IdEA—Developments with ESGF's system for Identity, Entitlement and Access Management ***Phil Kershaw (IS-ENES/BADC, philip.kershaw@stfc.ac.uk) and Rachana Ananthakrishnan (U. of Chicago, ranantha@uchicago.edu)***	We will present an update on progress with the activities identified in the roadmap for the development of the Earth System Grid Federation (ESGF) access control system—ESGF-IdEA—set out at this meeting last year. This defined a plan for enhancements and improvements to the system. Since then an ESGF-IdEA working team has been established and has been meeting to co-ordinate work. We will present developments including support for simplified browser-based single sign-on process and authentication with wget scripts without the need for X.509 certificates. We will also provide an assessment of the priorities for future work in the context of experience with the operational federation over the last year.
ESGF Transfer ***Eric Blau (U. of Chicago, blau@mcs.anl.gov), Rachana Ananthakrishnan (U. of Chicago, ranantha@uchicago.edu)***	During this session, we'll provide an update on the transfer capabilities available in ESGF. Updates include using latest GridFTP version of server, simplification of install process, "Science DMZ" friendly install options and improvements to wget for usability."
Automated Replication and Versioning ***Stephen Kindermann (IS-***	We will give an overview on past experiences made with replication and versioning and pressing issues. We will present some first concepts on how to tackle these issues, which involve the coherent assignment and use of persistent identifiers across the Earth System Grid Federation (ESGF) services. At the technical level, the first step is to embed resolvable identifiers in the

ENES2/DKRZ, kindermann@dkrz.de) and Tobias Weigel (IS-ENES2/DKRZ, weigel@dkrz.de)	netCDF headers of files submitted to ESGF. In a second phase, identical replicas could be identified and individual replica IDs stored in their metadata. This, however, only works if the necessary policies are enforced and services or tools are provided to encapsulate the whole functionality for daily operations. Individual data versions may also be made more accountable if persistent identifiers are assigned to each published set and if metadata is carried along that enabled the user interface to redirect users to the most recent version.
The ESGF Desktop: A Web–Desktop Interface to the ESGF Monitoring Infrastructure ***P. Nassisi (IS-ENES2/CMCC, paola.nassisi@cmcc.it), S. Fiore Aloisio (IS-ENES2/CMCC, sandro.fiore@unisalento.it) and G. Aloisio (IS-ENES2/CMCC, giovanni.aloisio@unisalento.it)***	The Earth System Grid Federation (ESGF) Desktop represents the graphical user interface of the ESGF monitoring infrastructure. It exploits the MVC design pattern and it relies on a strong adoption and implementation of Web 2.0 concepts such as mash-up, Google maps, and permalinks. It provides several views at different (hierarchical) granularity levels of the entire federation. In particular, through the ESGF Desktop, the user has the ability to visualize a set of statistics for both local (node-level) monitoring and global (institution-level and/or federation-level) monitoring. Real-time gadgets are also available. From an implementation point of view, the ESGF Desktop is a pure Javascript application with a set of RESTful APIs to make all of the metrics available to external applications. In terms of data usage statistics, the ESGF Desktop displays through specific gadgets: • The number of downloads; • The number of downloaded data sets; • The number of users that have downloaded some data; • The amount of data in term of downloaded gigabytes or terabytes; • The most downloaded (e.g. Top ten) data sets, variables and models; and • Data download client distribution. Additional gadgets are more related to multimedia content access (wiki pages, etc.) and desktop customizations.
Monitoring the Earth System Grid Federation Through the ESGF Dashboard ***P. Nassisi (IS-ENES2/CMCC, paola.nassisi@cmcc.it), S. Fiore Aloisio (IS-ENES2/CMCC, sandro.fiore@unisalento.it) and G. Aloisio (IS-ENES2/CMCC, giovanni.aloisio@unisalento.it)***	The Earth System Grid Federation (ESGF) Dashboard is a software component of the ESGF stack, responsible for collecting key information about the status of the federation in terms of: • *Network topology* (peer-groups composition) • *Node type* (host/services mapping) • *Registered users* (including their Identity Providers) • *System metrics* (e.g., round-trip time, service availability, CPU, memory, disk, processes, etc.) • *Real-time metrics* (e.g. RAM, CPU, etc.) • *Download statistics* (both at the Node and federation level) The last class of information is related to the data usage statistics, which are very important since they provide a strong insight of the Coupled Model Intercomparison Project Phase 5 experiment. During the presentation, the ESGF Dashboard architecture components (the information provider, the dashboard catalog, and the command line interface) will be presented jointly with a new scalable and configurable back-end storage model for long-term metrics.
Improved Usability and Support: esgf-swt ***Matthew Harris (DOE/LLNL, harris112@llnl.gov)***	Documentation: developers don't want to write it and users don't read it. This age-old argument has come to the forefront of Earth System Grid Federation discussions. With an aging and converted wiki, an up, down and then up again Askbot, and unachieved non-searchable email list, user support needs to be reconsidered and reorganized. The user must be able to define and support multiple software components to gain usage understanding. **Key points:** • What does the term user mean? • Who are our users? • Askbot beta runs. Now production • Creating a achieved email list and making it searchable • Wiki/Wikis standards, cleanup, uses
Revisiting the ESGF Node Manager for Federation Scalability ***Prashanth Dwarakanath (Linköping University, pchengi@nsc.liu.se) and Sasha Ames (DOE/LLNL, ames4@llnl.gov)***	The ESGF node manager is the component within the ESGF software stack that: • Gathers metrics • Shares node information across federated nodes • Facilitates user group management The present implementation of the node manager has scalability limitations arising from the peer-to-peer protocol. We present a design for next-generation node manager software and protocol that will address the exiting shortcomings of the current implementation and will offer additional features for evolving software components.

ESGF Metadata Search Evolution: esgf-mswt ***Luca Cinquini (NASA/JPL, Luca.Cinquini@jpl.nasa.gov)***	Arguably, one of the most important features of the Earth System Grid Federation (ESGF) is the capability to search, in real time, a system of metadata archives that are distributed around the world, and administered independently. Nonetheless, as the ESGF collaboration grows in scale and scope, its search capabilities must evolve to address new challenges, including: a) rigorous validation of metadata according to controlled vocabularies and schemas; b) partition of the search space according to project-specific criteria; c) dynamic configuration of peer circles; d) "big data" scalability; and e) new search operations such as temporal and geospatial constraints. This talk will discuss some ideas on how to upgrade the search infrastructure for the next generation projects, staring with the Coupled Model Intercomparison Project Phase 6.
Making the Case for the ESGF and Apache: Long-Term Software Stewardship ***Chris Mattmann (NASA/JPL, chris.a.mattmann@jpl.nasa.gov)***	In this talk I will give an overview of the Apache Software Foundation, its setup, meritocracy and governance structure. I will discuss Apache as a home for many long term and widely used software projects including HTTPD, Tomcat, and more recently de facto big data platforms like Spark, Mesos, Hadoop, Tika, Lucene and others. I will also identify Apache as a modern home for science-driven OSS, including the Apache OODT, and Open Climate Workbench projects. I will finally propose Apache as a potential home for the Earth System Grid Federation software stewardship and code base.

Session Seven: UV-CDAT Technical Development

Title and Presenter	Abstract
ESGF Compute Working Team (ESGF-CWT): Distributed Analytics Application Programming Interface (API) **(Session's Keynote)** ***Dan Duffy (NASA Center for Climate Simulation, daniel.q.duffy@nasa.gov)***	The model output from the Intergovernmental Panel on Climate Change (IPCC) Sixth Assessment Report (AR6) is estimated to create four to five times more data than is currently in the AR5 distributed archive. It is clear that data analysis capabilities currently available across the Earth System Grid Federation (ESGF) will be inadequate to allow for the necessary science to be done with AR6 data—the data will just be too big. A major paradigm shift from downloading data to local systems to perform data analytics must evolve to moving the analysis routines to the data and performing these computations on distributed platforms. In preparation for this need, the ESGF has started a Compute Working Team (CWT) to create solutions that allow users to perform distributed, high-performance data analytics on the AR6 data. The team will be designing and developing a general Application Programming Interface (API) to enable highly parallel, server-side processing throughout the ESGF data grid. This API will be integrated with multiple analysis and visualization tools, such as the Ultrascale Visualization Climate Data Analysis Tools (UV-CDAT), netCDF Operator, and others. This presentation will provide an update on the ESGF CWT's overall approach toward enabling the necessary storage proximal computational capabilities to study climate change using the AR6 extreme-scale distributed data archive. An update on the API will be provided along with a survey of the overall computational approaches being reviewed and studied by the members of the ESGF CWT.
ESGF Compute Node API ***Charles Doutriaux (DOE/LLNL, doutriaux1@llnl.gov)***	With each new round of the Coupled Model Intercomparison Project (CMIP)/Intergovernmental Panel on Climate Change, the volume of data served has grown about two orders of magnitudes. CMIP6 will be no exception and is expected to generate four to five times the amount of data than CMIP5. With such volumes it is not only impractical for any organization to hold all the data locally, but the end user will most likely not be able either to download locally the subset of data needed for his/her research both in terms of disk space and download bandwidth. One solution to this is to stop bringing the data to the scientists, but rather to bring their codes to the data. With this in mind, the Earth System Grid Federation (ESGF) compute working team was established. The primary charge to the Earth System Grid Federation Compute Working Team (esgf-cwt) is to allow ESGF users to execute analysis tools on high-end compute clusters, high-performance computers, cloud servers, and other forms of compute servers. In this talk we describe the state of the group, the decision taken so far, and the directions the groups is exploring.
Diagnostics: acme-dwt *Jeff Painter (DOE/LLNL, painter1@llnl.gov), Jim McEnerney (DOE/LLNL, mcenerney1@llnl.gov), and Brian Smith (DOE/ORNL, smithbe@ornl.gov)*	The Ultrascale Visualization Climate Data Analysis Tools (UV-CDAT) Diagnostics are one of the more important tools for climate scientists and code developers to compare model output with observations or another model. Data computed from model output or observations is plotted or sometimes tabulated. We have improved this diagnostic tool to offer twelve sets of plot for atmosphere and five for land; often with hundreds of possible plots per plot set. We can make plots for the UV-CDAT GUI, where they may be changed interactively; or we can write them as static image files for viewing with traditional tools such as web browsers. The diagnostic tools can

	be run from a GUI or with one of three command-line scripts.
Diagnostics and Web Informatics Support Infrastructure ***Brian Smith (DOE/ORNL, smithbe@ornl.gov)***	We have developed a number of tools to make viewing output produced by Ultrascale Visualization Climate Data Analysis Tools-based scripts easier for the end users. Part of that work has been an extensible Django-based backend and several front-end scripts to produce results similar to the NCAR diagnostics yet flexible enough to allow additional diagnostics and visual analysis tools to use them. In this talk, the "meta" diagostics master script and some of the Django back end components will be discussed.
ACME Exploratory Analysis and Diagnostics Viewer ***John Harney (ORNL/DOE, harneyjf@ornl.gov)***	The traditional Community Earth System Model diagnostics package provides scientists and modelers a way to quickly analyze the effectiveness of a climate model by computing climatological means of the large-scale simulations, and producing hundreds of plots and tables of the mean climate in a variety of formats. While the modeling community has successfully utilized this toolkit for a number of years, it is incompatible with modern workflows and methodologies. In this talk, we introduce the novel "Classic" diagnostics viewer, an improved interface for diagnostics. The Classic viewer is a key component of the Exploratory Analysis toolkit in the Accelerated Climate Model for Energy post-processing workflow, which utilizes the Ultrascale Visualization Climate Data Analysis Tools for efficient computation of key metrics and the Earth System Grid Federation for data archiving and cataloging. We will explore the various features offered by the classic viewer, as well as key features currently under development.
On Demand Data Reordering for Remote Data Processing and Exploration ***Timo Bremer (LLNL/DOE, bremer5@llnl.gov)***	The ViSUS client integrated into the Ultrascale Visualization Climate Data Analysis Tools allows interactive processing and visualization of large-scale ensembles of both local and remote data sets. However, it requires a data reordering to enable a progressive, out-of-core data access. In this talk we will introduce a new server infrastructure currently under development that provides a transparent re-ordering of data stored in an ESG node. This will enable any data severed by an ESG to be accessed through ViSUS and will allow an extensive processing and exploration of remote data.
ESGF Analysis and Visualization: Challenges and Opportunities ***Roland Schweitzer (NOAA/PMEL, roland.schweitzer@noaa.gov) and Kevin O'Brien (NOAA/PMEL, kevin.m.o'brien@noaa.gov)***	The Live Access Server from NOAA's Pacific Marine Environmental Laboratory is an independent web application that has a decades-long record of providing analysis and visualization of climate data. As such, we were fortunate to join the ESGF project during the run-up to the release of the Coupled Model Intercomparison Project (CMIP) Phase 5 data collection. During the initial planning, we developed an ambitious set of goals for a deeply integrated User Interface with a custom data set selection, region selection and other user interface (UI) controls as a direct part of the Gateway Node software stack. We further envisioned multiple backend engines, namely NCL, CDAT, and Ferret, producing products as desired by the installer of the system. However, having a custom and integrated UI proved to be too ambitious. In fact, to date, it seems the majority of the Earth System Grid Federation (ESGF) use has been focused on discovering data to download for local use. Offering services like those envisioned at the outset of the CMIP5 cycle requires a clear understanding of where and how the interaction with data set will take place. In some cases, that interaction takes place in environments that lack the flexibility for easy integration of authentication and authorization mechanisms. In this presentation, we will discuss opportunities that we envision to provide better visualization and subsetting capabilities in the ESGF context, while also recognizing the challenges that have prevented this from becoming a reality.
Web-based Visualization: Overview of Client and Server Side Techniques and their Use in GeoJS and CDATWeb ***Aashish Chaudhary (Kitware, aashish.chaudhary@kitware.com)***	With the power of Web 2.0 at the fingertips of the developers, it is now possible to create high-performance visualization for climate and geospatial domain experts in a web browser. Currently, this can be achieved in many ways. In this talk we will present various technologies that can provide practical solutions and how we are incorporating them in the development of CDATWeb, a web-based visualization framework for the Ultrascale Visualization Climate Data Analysis Tools.
Ultrascale Climate Data Visualization and Analysis Using UV-CDAT and DV3D ***Thomas Maxwell (NASA/GSFC, thomas.maxwell@nasa.gov) and Jerry Potter (NASA/GSFC, gerald.potter@nasa.gov)***	In collaboration with the Ultra-scale Visualization Climate Data Analysis Tools (UV-CDAT) development consortium, the National Aeronautics and Space Administration (NASA) National Center for Climate Simulation is developing climate data analysis and visualization tools for UV-CDAT. These tools feature workflow interfaces, interactive 3D data exploration, hyperwall and stereo visualization, automated provenance generation, parallel task execution, and streaming data parallel pipelines. NASA's DV3D is a UV-CDAT package that enables exploratory analysis of diverse and rich data sets from various sources including the Earth System Grid Federation (ESGF). DV3D's user-friendly interface places many complex visualization operations, previously sequestered within the exclusive domain of visualization specialists, at the fingertips of climate

	scientists. DV3D's tight integration into UV-CDAT seamlessly couples a wide range of high-performance climate data analysis operations with a rich palette of interactive visualization methods.
Workflow and Testing for Modern Software ***Aashish Chaudhary (Kitware, aashish.chaudhary@kitware.com)***	Software testing is an essential component in making sure that software meets a certain standard; open source or otherwise. To ensure the software quality for the modern software, it is important to setup a workflow that strengthens software testing for tools using a distributed code repository and code contribution from developers from various geo-locations. In this presentation, we will talk about modern software processes and testing for desktop- and web-based projects.
GIS Capabilities in UV-CDAT ***Ben Koziol (NOAA/ESRL, ben.koziol@noaa.gov)***	Following the Earth System Grid Federation/Ultra-scale Visualization Climate Data Analysis Tools (UV-CDAT) Face-to-Face meeting in December 2013, a plan was initiated to bring low-level geographic information system capabilities into UV-CDAT (i.e. subsetting via ESRI Shapefile) through functionality provided by OpenClimateGIS (OCGIS). OCGIS is an open source Python package designed for geospatial manipulation, subsetting, computation, and translation of climate data sets stored in local NetCDF files or files served through OpenDAP data servers. In addition to an overview of OCGIS, this presentation will provide a description of the proposed integration plan with UV-CDAT and discuss the current state of development. Furthermore, an overview of OCGIS development with ESMPy (the Python interface to the Earth System Modeling Framework [ESMF]) will be described. The ESMPy-OCGIS connection is an additional pathway for bringing GIS-like operations into UV-CDAT. ESMP, the previous version of the ESMF Python interface, is currently used by UV-CDAT for regridding operations.

Day 3: Thursday, 11 December 2013
ESGF and UV-CDAT Technical Presentations and Discussions
Technical Interoperability Discussions

Session Eight: Other Related Community Contributions

Title and Presenter	Abstract
ES-DOC **(Set up Go-To-Meeting)** ***Mark Greenslade (IS-ENES/IPSL, momipsl@ipsl.jussieu.fr)***	During 2014, the ES-DOC project (http://es-doc.org) extended its ecosystem by deploying significant upgrades of its documentation creation, search, viewing, and comparison tools. The ES-DOC questionnaire and the pyesdoc scripting library will serve as the pathway for generating Coupled Model Intercomparison Project Phase 6 documentation. Underpinning these tools is a publically available web service Application Programming Interface in support of documentation search, publication, and comparison. ES-DOC continues to support and leverage the emergent Metafor CIM documentation standard and is proactively assisting other projects to do likewise. Furthermore, the project has improved its own software development process via streamlined deployments and deepened automated testing.
Towards a Controlled Vocabulary Service **(Set up Go-To-Meeting)** ***Mark Greenslade (IS-ENES/IPSL, momipsl@ipsl.jussieu.fr)***	DKRZ hosts the Intergovernmental Panel on Climate Change Data Distribution Center, which provides long-term access to Coupled Model Intercomparison Project (CMIP) data for the interdisciplinary (re-)use. Beyond permanent and persistent data access, the center needs to provide detailed documentations, a uniform data quality, and DataCite DOI data citations in order to enable data users to accept or even trust the data and to give credit to data creators.
System for Offline Data Access (SODA) ***Prashanth Dwarakanath (Linköping University, pchengi@nsc.liu.se)***	The System for Offline Data Access (SODA) is one of the activities being carried out under the CLIPC (Climate Information Platform for Copernicus) FP7 EU project, under the task "Dynamic Tape-archive extraction and post-processing." The deliverable of this exercise is a generic system that allows ESGF users to search and retrieve data stored on different kinds of offline tape systems. This work targets the publication of data from the EURO4M project present on the MARS system at SMHI-LIU as a demonstrator.
Climate and Forecast (CF) Conventions ***Karl Taylor (DOE/LLNL,***	The CF Conventions define a standard structure for NetCDF files. They are widely followed in the climate community because they provide the machine-readable semantics that software needs to process climate data. Most of the data on ESGF follows the conventions, and the requirements for Coupled Model Intercomparison Project Phases 3–6 build on the CF Conventions. In the last

taylor13@llnl.gov), Jeff Painter (DOE/LLNL, painter1@llnl.gov), Matthew Harris (DOE/LLNL, harris112@llnl.gov)	year we have updated the CF Conventions document with some of the settled issues in our Trac issue-tracker system. We will make more such changes soon and then release it as CF Conventions version 1.7. Then CF Conventions 1.8 will contain a new chapter, describing the Gridspec specification, a structured way to describe unstructured grids. Most of our effort in the last year has been devoted to dealing with a hardware failure in the web server. We built a new github-based web server for the CF Conventions site, and moved the Trac system first to one new server and then another. As these new hosts grow more stable we will be able to shift back to work on the document itself.
Preparing CMOR for CMIP6 *Charles Doutriaux (DOE/LLNL, doutriaux1@llnl.gov) and Karl Taylor (DOE/LLNL, taylor13@llnl.gov)*	One of the key components to the success of the Model Intercomparison Projects, and in particular the Coupled Model Intercomparison Project (CMIP), has been the availability of multiple models output generated using common standard, which makes them easy to analyze and compare. CMIP5 came with a full set of documents describing in details which variables should be stored and how they should be made available back to the community, be it format, naming conventions, description conventions, etc. In addition to these many documents, CMIP5 provided a multi-language tool to help scientists producing conforming data. This tool is the Climate Model Output Rewriter (CMOR). This talk describes the state of the CMOR software and what steps are envisioned to get it ready for CMIP6. In particular we will look at what needs to be added to conform to the new CMIP6 requirement and also to adapt to data that are not necessarily issued from the modeling community (observation, reanalysis, etc.).
Ophidia: A Big Data Analytics Framework for eScience *G. Aloisio (IS-ENES2/Univ. of Italy, giovanni.aloisio@unisalento.it), S. Fiore (IS-ENES2/CMCC, sandro.fiore@unisalento.it)*	The Ophidia project is a research effort on big data analytics–facing scientific data analysis challenges in the climate change domain. It provides parallel (server-side) data analysis, an internal storage model, and a hierarchical data organization to manage large amount of multidimensional scientific data. The Ophidia analytics platform provides several data operators to manipulate multidimensional data sets. Some relevant examples include: 1) data sub-setting (slicing and dicing), 2) data aggregation, and 3) data analysis. Additionally, the Ophidia framework provides about 100 primitives to perform time series analysis, sub-setting, and data aggregation on large arrays of scientific data. Multiple primitives can be also nested to implement a single more complex task (e.g., aggregating by sum a subset of the entire array). The entire Ophidia software stack has been deployed at CMCC on 24 nodes (16-cores/node) of the Athena HPC cluster. A comprehensive benchmark and test cases are being defined with climate scientists to extensively test all of the features provided by the system. Preliminary experimental results are already available and have been published on scientific research papers.
JASMIN: Cloud Computing System *Phil Kershaw (IS-ENES/BADC, philip.kershaw@stfc.ac.uk)*	Cloud computing is a disruptive technology that presents some significant opportunities and challenges for successful exploitation by the scientific community. It promises the ability to provide near-limitless computing resources on demand, pooling of resources between different groups and activities, and broad network access to support widely distributed communities of users. However, the specialist requirements for scientific workloads can be at odds with the commodity-based infrastructure typically provided by public clouds. One solution is to deploy a private cloud customized to meet these needs, but in doing so there are technical and operational challenges to be tackled. In the case of climate science and earth observation applications, the increasing volumes of data mean that access to large capacity storage with performant input/output (IO) to compute is a critical factor for effective processing and analysis. In this presentation we will explore experiences tailoring cloud computing technology and service models for a private cloud for JASMIN. JASMIN is a large storage and analysis facility for the UK climate science community and its international partners. It was first established two years ago and funded through the UK Natural Environment Research Council (NERC). Over the past year, a second phase of development has been underway expanding the system to over 3,000 processing cores and around 12 PB of disk-based storage. In addition the scope of the service has been broadened to the big data analysis needs of the entire UK environmental science community. JASMIN in its first phase rapidly demonstrated the benefits of its global file system and high performance IO between storage and compute. One of the challenges in this new phase is to broaden access to the resource to better meet the needs of the so-called long tail of science. This area is being addressed directly with some of the first tenancies on the new cloud service. Among these, a service hosting Linux desktop environments has been customized with applications and libraries for an individual community's needs. Similarly, the popular IPython Notebook application was hosted. In each case users obtain access to interfaces and tools they are familiar with but underpinned by the extensive compute and storage resources of a large centralized facility. This presentation will explore the technical and organizational aspects involved in developing and deploying infrastructure for an operational cloud service.

ESGF in an OpenStack Cloud ***Ben Evans (ANU/NCI, Ben.Evans@anu.edu.au)***	Cloud systems provide a flexible platform for supporting a range of complex services that can easily scale in resources, be managed and upgraded, and is able to be linked with other inter-operating functionality and as the capabilities are developed and released. The Earth System Grid Federation (ESGF) is a typical example of a software environment that is well suited to this. In this talk I will discuss some of the key components of our installation, and some of the infrastructure issues that have been addressed, and issues that need to be addressed in future deployment of the ESGF system.
Requirements for a Biology node on ESGF ***Patrik D'haeseleer (DOE/LLNL, dhaeseleer2@llnl.gov) and Sasha Ames (DOE/LLNL, ames4@llnl.gov)***	The federated database platform provided by the Earth System Grid Federation (ESGF) can be of use in other disciplines as well, especially in biology, where there are hundreds of existing databases for specific purposes, and a chronic need for more integration and interoperability between them. One key opportunity lies in incorporating epidemiological data on ESGF. Disease outbreaks are inherently spatiotemporal, so this type of data should mesh well with the existing climate modeling infrastructure. Some diseases are known to exhibit seasonal, weather and climate variations, so it should be possible to predict the likelihood of disease outbreak based on weather patterns, how the endemic range of diseases such as dengue is expected to shift with climate change, and perhaps even which mutations in pathogens are associated with climatic adaptations. We will focus on two key types of data: epidemiological data describing disease outbreaks over time in different locations, and sequence data, which is one of the most fundamental data types for modern-day biology.